



THE UNIVERSITY
of ADELAIDE

Anomaly Detection in Computer Vision and Medical Imaging

by

Yu Tian

A thesis submitted for the degree of

Doctor of Philosophy

June 15, 2022

Australian Institute for Machine Learning (AIML)

The University of Adelaide

Contents

Declaration	xxiii
Acknowledgements	xxv
Publications	xxvii
Abstract	xxxix
1 Introduction	1
1.1 Anomaly Detection Setups	1
1.2 Motivation	3
1.3 Contributions and Thesis Outline	6
2 Literature Review	11
2.1 Unsupervised Anomaly Detection	11
2.2 Weakly Supervised and Few-shot Anomaly Detection	13
2.3 Pixel-wise Anomaly Detection in Semantic Segmentation	14
2.4 Anomaly Detection Datasets	15
3 Photoshopping Colonoscopy Frames	21
3.1 Introduction	21
3.2 Related work	24
3.3 Data Set and Methods	25
3.3.1 Data Set	25
3.3.2 Methods	26
3.4 Experiment	27
3.4.1 Experimental Setup	27
3.4.2 Anomaly Detection Results	28
3.4.3 Image Reconstruction from ADGAN	28
3.5 Conclusions	28

4	Constrained Contrastive Distribution Learning for Unsupervised Anomaly Detection and Localisation in Medical Images	33
4.1	Introduction	34
4.2	Method	35
4.2.1	Constrained Contrastive Distribution Learning	35
4.2.2	Anomaly Detection and Localisation	37
4.3	Experiments	38
4.3.1	Dataset	38
4.3.2	Implementation Details	39
4.3.3	Ablation Study	40
4.3.4	Comparison to SOTA Models	41
4.4	Conclusion	43
5	Self-supervised Pseudo Multi-class Pre-training for Unsupervised Anomaly Detection and Segmentation in Medical Images	47
5.1	Introduction and Background	48
5.2	Related Work	50
5.3	Method	51
5.3.1	PMSACL Pre-training	52
5.3.2	MedMix Augmentation	54
5.3.3	Anomaly Detection and Segmentation	55
5.4	Experiments	56
5.4.1	Datasets	56
5.4.2	Implementation Details	57
5.4.3	Evaluation Measures	58
5.4.4	Anomaly Detection Results	58
5.4.5	Anomaly Localisation Results	63
5.4.6	Qualitative Results	64
5.4.7	Ablation Study	66
5.5	Conclusion	69
6	Deep One-Class Classification via Interpolated Gaussian Descriptor	73
6.1	Introduction	73
6.2	Related Work	75
6.3	Method	76
6.3.1	Interpolated Gaussian Descriptor (IGD)	77
6.3.2	Theoretical Guarantees	80
6.3.3	Training and Inference	82
6.4	Experiments	84
6.4.1	Datasets and Evaluation Metric	84
6.4.2	Implementation Details	84

6.4.3	Experiments on MNIST, Fashion MNIST and CIFAR10	86
6.4.4	Experiments on MVTec AD	87
6.4.5	Experiments on Medical Datasets	88
6.4.6	Visualisation of the Distribution of Testing Samples	88
6.4.7	Ablation Study	88
6.4.8	Experiments on Small/Contaminated Training Sets	89
6.5	Discussion	91
6.6	Conclusion	92
7	Unsupervised Anomaly Detection in Medical Images with a Memory-augmented Multi-level Cross-attention Masked Autoencoder	95
7.1	Introduction	96
7.2	Method	97
7.2.1	Memory-augmented Multi-level Cross-attention Masked Autoencoder (MemMC-MAE)	97
7.2.2	Anomaly Detection and Segmentation	100
7.3	Experiments and Results	101
7.4	Conclusion	104
8	Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning	107
8.1	Introduction	108
8.2	Related Work	110
8.3	The Proposed Method: RTFM	111
8.3.1	Theoretical Motivation of RTFM	112
8.3.2	Multi-scale Temporal Feature Learning	114
8.3.3	Feature Magnitude Learning	116
8.3.4	RTFM-enabled Snippet Classifier Learning	116
8.4	Experiments	117
8.4.1	Data Sets and Evaluation Measure	117
8.4.2	Implementation Details	118
8.4.3	Results on ShanghaiTech	118
8.4.4	Results on UCF-Crime	119
8.4.5	Results on XD-Violence	119
8.4.6	Results on UCSD-Peds	121
8.4.7	Sample Efficiency Analysis	121
8.4.8	Subtle Anomaly Discriminability	122
8.5	Computational Efficiency	123
8.5.1	Ablation Studies	123
8.5.2	Qualitative Analysis	124
8.6	Conclusion	125

9	Contrastive Transformer-based Multiple Instance Learning for Weakly Supervised Polyp Frame Detection	129
9.1	Introduction and Background	129
9.2	Method	131
9.2.1	Convolutional Transformer MIL Network	132
9.2.2	Transformer-based MIL Training	132
9.3	Experiments and Results	135
9.3.1	Dataset	135
9.3.2	Implementation Details	135
9.3.3	Evaluation on Polyp Frame Detection	136
9.3.4	Ablation Study	137
9.4	Conclusion	137
10	Few-Shot Anomaly Detection for Polyp Frames from Colonoscopy	141
10.1	Introduction	141
10.2	Related Work	143
10.3	Data Set and Method	144
10.3.1	Dataset	144
10.3.2	Method	145
10.4	Experiment	147
10.4.1	Experimental Setup	147
10.4.2	Anomaly Detection Results	147
10.5	Conclusion	150
11	Pixel-wise Energy-biased Abstention Learning for Anomaly Segmentation on Complex Urban Driving Scenes	153
11.1	Introduction	153
11.2	Related work	156
11.3	Method	158
11.3.1	Training Set	158
11.3.2	Pixel-wise Energy-biased Abstention Learning (PEBAL)	158
11.3.3	Training and Inference	160
11.4	Experiment	161
11.4.1	Datasets	161
11.4.2	Implementation Details	162
11.4.3	Evaluation Measures	162
11.4.4	Comparison on Anomaly Segmentation Benchmarks	162
11.4.5	Ablation Study	166
11.4.6	Outlier Samples, Calibration and Efficiency	167
11.5	Conclusions and Discussions	168

12 Conclusion	169
12.1 Limitations and Future Work	171
A IGD (Chapter 6) Appendix	173
A.1 Datasets	173
A.2 Global and Local IGD Models	173
A.3 Multi-scale Structure Similarity Index (MS-SSIM) Score	174
A.4 Implementation Details	175
A.5 Visualisation of the Distribution of Testing Samples	175
A.6 Correctness Proof	176
A.7 Convergence Conditions Proof	177
A.8 Class-level Results	178
A.9 Qualitative Localisation Results	179
B RTFM (Chapter 8) Appendix	183
B.1 Theoretical Motivation of RTFM	183
B.2 Computational Efficiency	184
B.3 Temporal Dependency	184
B.4 Ablations for k and m	184
C PEBAL (Chapter 11) Appendix	187
C.1 Qualitative results	187
C.2 More AUC results	187
C.3 Hyper-parameters Selection	188
C.4 Training Details on Cityscapes	189
C.5 Results Based on Different DeepLabv3+ Checkpoint	189
Bibliography	191

List of Tables

1.1	Thesis contributions: We propose different deep learning methods for various anomaly detection tasks, including unsupervised anomaly detection without abnormal training data, weakly supervised anomaly detection with weakly labelled video training data, and few-shot anomaly detection with few labelled image data.	7
3.1	Comparison between our proposed ADGAN and other state of the art methods.	28
4.1	Ablation study of the loss terms in (4.1) on Hyper-Kvasir, using IGD as anomaly detector.	41
4.2	Anomaly localisation: Mean IoU results on Hyper-Kvasir on 5 different groups of 100 images with ground truth masks. * indicates that we pretrained the geometric transformation-based anomaly detection [77] using IGD [34] as the UAD method.	41
4.3	Anomaly detection: AUC results on Hyper-Kvasir, Liu et al.’s colonoscopy and LAG, respectively. * indicates that the model does not use Imagenet pre-training.	42
5.1	Anomaly detection testing results on Hyper-Kvasir in terms of AUC, Specificity, Sensitivity and Accuracy. Best results are highlighted.	59
5.2	Anomaly detection testing results on LAG in terms of AUC, Specificity, Sensitivity, Precision and Recall. Best results are highlighted. . .	60
5.3	Anomaly detection testing results on Liu et al.’s colonoscopy in terms of AUC, Specificity, Sensitivity and Accuracy. * indicates that the model does not use ImageNet pre-training. Best results are highlighted.	61
5.4	Anomaly detection testing results on Covid-X in terms of AUC, Specificity, Sensitivity and Accuracy, respectively. Best results are highlighted.	62

5.5	The standard deviation of five-run experimental results on the Hyper-Kvasir, LAG and Covid-X based on the PMSACL pre-trained PaDiM anomaly detector. This results should be studied together with the results shown in Tables 5.1, 5.2, 5.4.	62
5.6	Anomaly localisation: Mean IoU, Dice and PRO-AUC testing results on Hyper-Kvasir on 5 different groups of 100 images with ground truth masks. Best results for each case are highlighted.	64
5.7	Anomaly localisation: Mean IoU, Dice and Pro-AUC testing results on abnormal samples from LAG test set. Best results are highlighted.	64
5.8	Ablation study of the PMSACL loss terms on the test sets of Hyper-Kvasir and LAG, using PaDiM [43] as anomaly detector, with our MedMix as strong augmentations.	67
5.9	Ablation studies with different self-supervised pre-training approaches on Hyper-Kvasir testing set. PaDiM [43] is used as the anomaly detector. Best results are highlighted.	68
6.1	Anomaly detection: mean AUC testing results on MNIST, CIFAR10 and Fashion MNIST. The results are split into 'Scratch' (without any pre-training), pretrained with 'ImageNet', and self-supervised learning ('SSL'). Bold numbers represent the best result (within 0.5%) for each data set, discriminated by Scratch, SSL or ImageNet.	85
6.2	Anomaly detection: mean testing accuracy and AUC on MVTec AD produced by the SOTA and our IGD.	87
6.3	Anomaly localisation: mean pixel-level AUC testing results on the anomalous images of MVTec AD.	88
6.4	Anomaly detection: AUC testing results on two medical datasets: Hyper-Kvasir and LAG.	89
6.5	Ablation study of our method on CIFAR10 using anomaly detection mean testing AUC w.r.t standard OCC setup (AUC - Full), small training set containing 20% of training data (AUC - ST), and anomaly contaminated training set with 10% contamination (i.e., 10% of the anomalous samples are removed from the testing set and inserted into the training set) (AUC - AC). MSE denotes the baseline deep autoencoder with MSE loss, REC denotes the baseline deep autoencoder with MS-SSIM + MAE losses, GAC denotes our proposed Gaussian anomaly classifier, INTER represents our interpolation regularisation. The encoder of all above methods are initialised based on the knowledge distillation from ImageNet.	90
6.6	Mean testing AUCs on CIFAR10 and MVTec with small training sets, where REC=MS-SSIM+MAE losses.	90

6.7	Mean testing AUCs on CIFAR10 and MVTEC with different contamination noise rates. REC defined in Tab. 6.6.	91
7.1	Anomaly detection AUC test results on Covid-X and Hyper-Kvasir. CCD+IGD* [221] requires at least 2×longer training time than other approaches in the table because of a two-stage self-supervised pre-training and fine-tuning.	100
7.2	Ablation study on Covid-X of the encoder’s memory-augmented operator (Mem-Enc) and the decoder’s multi-level cross-attention (MC-Dec).	103
7.3	Anomaly localisation: Mean IoU test results on Hyper-Kvasir on 5 groups of 100 images.	103
8.1	Comparison of frame-level AUC performance with other SOTA un/weakly-supervised methods on ShanghaiTech. * indicates we retrain the method in [211] using I3D features. Best result in red and second best in blue	119
8.2	Frame-level AUC performance on UCF-Crime. * indicates we retrain the method in [211] using I3D features. Best result in red and second best in blue	120
8.3	Comparison of AP performance with other SOTA un/weakly-supervised methods on XD-Violence. * indicates we retrain the method in [211] using I3D features. Best result in red and second best in blue	120
8.4	Comparison of AUC performance with other SOTA weakly-supervised methods on UCSD Ped2. * indicates we retrain the method in [211] using I3D features. Best result in red and second best in blue	121
8.5	Ablation studies of our method on ShanghaiTech and UCF-Crime.	124
9.1	Comparison of frame-level AUC and AP performance with other SOTA WVADs on colonoscopy dataset using the same I3D feature extractor.	136
9.2	Ablation studies for polyp frame detection. The linear network with top-k MIL ranking loss is considered as the baseline, and CTE denotes the Convolutional Transformer Encoder.	137
10.1	Comparison between our proposed FSAD-NET and other state of the art zero-shot and few-shot anomaly detection methods.	148
11.1	Anomaly segmentation results on LostAndFound testing set, with WideResnet34 backbone. All methods use the same segmentation models. * indicate that the model requires additional learnable parameters. † indicates that the results are obtained from the official code with our WideResnet34 backbone.	163

11.2	Comparison with previous approaches on Fishyscapes Leaderboard . We achieve a new state-of-the-art performance among the approaches that require extra OoD data, and without re-training the segmentation networks and extra networks on Fishyscapes Leaderboard.	164
11.3	Anomaly segmentation results on Fishyscapes validation sets (LostAndFound and Static), and the Road Anomaly testing set , with WideResnet34 backbone. * indicate that the model requires additional learnable parameters. † indicates that the results are obtained from the official code with our WideResnet34 backbone. Best and second best results in bold.	164
11.4	Anomaly segmentation results on Fishyscapes validation sets (LostAndFound and Static), and the Road Anomaly testing set , with Resnet101 backbone. * indicate that the model requires additional learnable parameters. † indicates that the results are obtained from the official code with our Resnet101 backbone. Best and second best results in bold.	165
11.5	Ablation studies for anomaly segmentation on LostAndFound , with WideResnet34 backbone, where all proposed modules are trained with COCO OE images with AnomalyMix. CE denotes the baseline method that adds an extra OoD class to learn the OE training samples with cross-entropy (first row).	166
11.6	The performance comparison of our approach on Fishyscapes benchmark w.r.t different diversity of OE classes (mean results over six random seeds), in terms of AP and FPR95.	166
11.7	The performance comparison of our approach on Fishyscapes benchmark w.r.t different amount of OE training samples (mean results over six random seeds), in terms of AP and FPR95.	168
A.1	Anomaly detection : mean testing accuracy and AUC on MVTec AD produced by the SOTA and our method.	176
A.2	Anomaly detection : class-level testing AUC on MNIST produced by the SOTA and our methods.	179
A.3	Anomaly detection : class-level testing AUC on FMNIST produced by our methods.	180
A.4	Anomaly detection : class-level testing AUC on CIFAR10 produced by the SOTA and our methods.	180
A.5	Anomaly localisation : class-level testing pixel-wise localisation AUC results on the anomalous images of MVTec AD produced by our methods.	181
C.1	AUC testing results (mean results over six random seeds) of our approach on Fishyscapes benchmark w.r.t. different diversity of OE classes	187

C.2	AUC testing results (mean results over six random seeds) of our approach on Fishyscapes benchmark w.r.t. different amount of OE training samples	189
C.3	Anomaly segmentation results on Fishyscapes validation sets (LostAndFound and Static), and the Road Anomaly testing set , with WideResnet34 backbone under cv0 standard train/val split.	190

List of Figures

1.1	Taxonomy of three types of deep anomaly detection models explored in this thesis.	3
1.2	Anomaly detection applications explored in this thesis: road obstacles, industrial defect detection, Covid-19 detection from Chest X-ray, polyp detection from colonoscopy frames, glaucoma detection from fundus screening images, and violence detection from surveillance videos.	6
3.1	Top row shows test images containing polyps (highlighted with a red ellipse), which are considered to be anomalies in our framework. Bottom row shows the reconstructed images by our ADGAN model, which deviate with their top row input images leading to high reconstruction errors. Note that given that the ADGAN model was trained with images without polyps, it is biased to reconstruct images without polyps, as clearly seen in these examples.	22
3.2	Our proposed ADGAN model trains the visual generator, visual discriminator, latent generator and latent discriminator using adversarial training (left). During testing, the input image is processed by the latent generator and the produced latent embedding is used by the visual generator to produce the output image, which is then compared with the input image to compute the anomaly score.	25
4.1	Our proposed CCD framework. Left shows the proposed pre-training method that unifies a contrastive distribution learning and pretext learning on both global and local perspectives (Sec. 4.2.1), Right shows the inference for detection and localisation (Sec. 4.2.2).	36
4.2	Left: Anomaly detection performance results based on different batch sizes of self-supervised pre-training. Right: Anomaly detection performance in terms of different types of strong augmentations. Both results are on Hyper-Kvasir test set using IGD as anomaly detector.	40
4.3	Qualitative results of our localisation network based on IGD with self-supervised pre-training on the abnormal images from Hyper Kvasir [21] test set.	42

5.1	<p>PMSACL: our proposed self-supervised pre-training for UAD trains four classes of images: the normal images formed by the weak augmentations in distribution \mathcal{A}_0 (blue markers) and three classes of synthesised abnormal images formed by the strong augmentation in distributions $\{\mathcal{A}_n\}_{n=1}^3$ (green, pink and orange markers). The optimisation uses a constrained contrastive learning that trains a four-class classification problem. The different types of strong augmentations are produced by MedMix that introduces a varying number of fake lesions by cutting patches from the normal training images, altering them with random color jittering, Gaussian noise and non-linear intensity transformations, and pasting them to other normal training images.</p>	51
5.2	<p>Examples of our MedMix data augmentation, showing augmentation \mathcal{A}_0 containing zero synthetic anomalies (leftmost column) and increasingly stronger augmentations $\{\mathcal{A}_n\}_{n=1}^3$ (second to fourth columns) with different number of synthetic anomalies (from one to three).</p>	55
5.3	<p>Localisation of four abnormal images from Hyper Kvasir [121], with their predictions (Pred) and ground truth annotations (GT), using PaDiM with PMSACL pre-training.</p>	63
5.4	<p>Localisation of four abnormal images from LAG [121], with their predictions (Pred) and ground truth attention maps (GT), using IGD with PMSACL pre-training.</p>	63
5.5	<p>Visual detection results and anomaly scores produced by the PMSACL pre-trained IGD on three different datasets: Hyper-Kvasir (top), LAG (middle), Covid-X (bottom). Anomaly scores > 0.5 classifies the image as positive, otherwise, the image is classified as negative. Correctly classified images are marked with green boxes, and incorrectly classified cases are marked with red boxes.</p>	65
5.6	<p>t-SNE results of the image representations of the test set of Hyper-Kvasir [21] learned by IGD [34] after being pre-trained on ImageNet [47], or self-supervised with DROC [208], CCD [221], and our PMSACL. Compared to other methods, PMSACL clusters the normal image representations (blue points) in a tighter and denser region, and separates anomalous representations into three clusters (red points), which can be associated with the three classes of synthesised abnormal images formed by simulating a varying number of lesions of different sizes and appearance in the normal images.</p>	66

5.7	Anomaly detection testing results in terms of different types of strong augmentations (i.e., Cutmix, Gaussian noise, Rotation, Permutation, and our MedMix) on Hyper-Kvasir and Covid-X, where our PMSACL is used as self-supervised pre-training, and IGD [34] is used as the anomaly detector.	67
5.8	Influence of the number of MedMix augmentation distributions $ \mathcal{A} $ in (11.2) on the AUC results of Hyper-Kvasir testing set, where PaDiM [43] is used as the anomaly detector.	68
6.1	Mean testing AUC of DSVDD [191], and our proposed IGD trained with the CIFAR10 training set contaminated with 1%, 5% and 10% of anomalous samples (left), and small training sets, consisting of 20%, 60%, and 100% of the CIFAR10 training set (right).	74
6.2	Our IGD consists of an encoder that transforms image \mathbf{x} into representation \mathbf{z} , a decoder to reconstruct the image (trained with MS-SSIM and MAE losses), a Gaussian anomaly classifier trained to push the normal image representation close to the centre of the estimated normal image distribution (denoted by a Gaussian with mean μ and standard deviation σ), and a critic module that constrains the likelihood maximisation by predicting the interpolation coefficient α that produces a convex combination of training sample representations. Note that critic is a module similar to a GAN discriminator.	76
6.3	Example of the multi-scale structural and non-structural anomaly localisation result for an MVTEC AD [13] image, using both the local and global IGD models. The global model tends to produce smooth results but with some mistakes, while the local model produces jagged results, but without the global mistakes, so by combining the two results, we obtain a smooth and correct anomaly heatmap.	82
6.4	Qualitative results of our anomaly localisation results on the MVTEC AD (red = high probability of anomaly). Top, middle and bottom rows show the testing images, ground-truth masks and predicted heatmaps, respectively.	86
6.5	t-sne visualisation from MVTEC (class bottle).	89

7.1	Top: overall MemMC-MAE framework. Yellow tokens indicate the unmasked visible patches, and blue tokens indicate the masked patches. Our memory-augmented transformer encoder only accepts the visible patches/tokens as input, and its output tokens are combined with dummy masked patches/tokens for the missing pixel reconstruction using our proposed multi-level cross-attentional transformer decoder. Bottom-left: proposed memory-augmented self-attention operator for the transformer encoder, and bottom-right: proposed multi-level cross-attention operator for the transformer decoder.	98
7.2	Segmentation results of our proposed method on Hyper-Kvasir [21], with our predictions (Pred) and ground truth annotations (GT).	102
7.3	Reconstruction of testing images from Covid-X (Top) and Hyper-Kvasir (Bottom). For each triplet, we show the masked image (left), our MemMC-MAE reconstruction (middle), and the ground-truth (right). Normal testing images are marked with green boxes, and anomalous ones are marked with red boxes.	102
8.1	RTFM trains a feature magnitude learning function to improve the robustness of MIL approaches to normal snippets from abnormal videos, and detect abnormal snippets more effectively. Left: temporal feature magnitudes of abnormal and normal snippets ($\ \mathbf{x}^+\ $ and $\ \mathbf{x}^-\ $), from abnormal and normal videos (\mathbf{X}^+ and \mathbf{X}^-). Assuming that $\mu = 3$ denotes the number of abnormal snippets in the anomaly video, we can maximise the $\Delta\text{score}(\mathbf{X}^+, \mathbf{X}^-)$, which measures the difference between the scores of abnormal and normal videos, by selecting the top $k \leq \mu$ snippets with the largest temporal feature magnitude (the scores are computed with the mean of magnitudes of the top k snippets). Right: the $\Delta\text{score}(\mathbf{X}^+, \mathbf{X}^-)$ increases with $k \in [1, \mu]$ and then decreases for $k > \mu$, showing evidence that our proposed RTFM-enabled MIL model provides a better separation between abnormal and normal videos when $k \approx \mu$, even if there are a few normal snippets with large feature magnitudes.	108
8.2	Our proposed RTFM receives a $T \times D$ feature matrix \mathbf{F} extracted from a video containing T snippets. Then, MTN captures the long and short-range temporal dependencies between snippet features to produce $\mathbf{X} = s_\theta(\mathbf{F})$. Next, we maximise the separability between abnormal and normal video features and train a snippet classifier using the top- k largest magnitude feature snippets from abnormal and normal videos.	114

8.3	Our proposed MTN consists of two modules. The module on the left uses the pyramid dilated convolutions to capture the local consecutive snippets dependency over different temporal scales. The module on the right relies on a self-attention network to compute the global temporal correlations. The features from the two modules are concatenated to produce the MTN output.	115
8.4	Anomaly scores and feature magnitude values of our method on UCF-Crime (<i>stealing079,shoplifting028, robbery050 normal876</i>), and ShanghaiTech (<i>01_0052, 01_0053</i>) test videos. Pink areas indicate the manually labelled abnormal events.	121
8.5	AUC w.r.t. the number of abnormal training videos.	122
8.6	AUC results w.r.t. individual classes on UCF-Crime.	123
9.1	(a) The architecture of our method consists an I3D [24] snippet feature extractor and a Convolutional Transformer MIL Network. The I3D features are considered as snippet feature tokens to the transformer to predict snippet-wise anomaly scores using a snippet classifier. The Cls token is applied for a video classifier to predict if a video contains anomalies. The output features from the transformer are utilised to mine hard and easy snippets from normal and abnormal videos. The anomaly scores and hard/easy snippet representations are optimised by three proposed losses in (9.1). (b) The proposed Temporal Convolutional Transformer Layer replaces the linear projection with depthwise separable convolution (DW Conv1D) [38].	132
9.2	Hard abnormal snippet mining algorithm to select temporal edge snippets and missed pseudo abnormal snippets. Those two types of hard anomalies represent: 1) transitional frames where polyps may be partially visible, or 2) subtle (i.e., small and flat) polyps that can lead to incorrect low anomaly scores.	134
9.3	Anomaly scores (orange curve) of our method on test videos. Pink areas indicate the labelled testing abnormal events.	137
10.1	Depiction of the three different approaches to handle few-shot and zero-shot anomaly detection. Our proposed FSAD-NET demonstrate better deviations between normal and abnormal samples	142
10.2	The first stage of FSAD-NET training consists of modelling the encoder by maximising the MI between normal training images and embeddings in a global and local manner and by minimising the divergence of embeddings and a prior distribution [98]. The embeddings produced by the encoder are then used to train the SIN using a contrastive-like loss [167].	145

10.3	AUC mean and standard deviation of FSAD-NET computed over different number of abnormal training images.	149
10.4	True positive (TP), true negative (TN), false positive (FP) and false negative(FN) results produce by FSAD-NET (Negative = frame with polyp).	149
11.1	Anomaly segmentation overview. From the input image (anomaly highlighted with a yellow box), the initial prediction shows the original segmentation results with anomalies classified as a one of the pre-defined inlier classes. Anomaly predictions by the previous SOTA Meta-OoD [27] and our method show an anomaly map with high scores (in yellow and red) for anomalous pixels, where our approach shows less false positive and false negative detections. Consequently, our method can detect small and distant anomalies (row 2) and blurry/unclear anomalies (rows 1, 3, 4) more accurately than Meta-OoD [27]. In our final prediction , anomalous pixels are coloured as cyan. Some anomalies are small and blurred (e.g., row 2), so please zoom in the PDF for better visualisation.	155
11.2	PEBAL. The pixel-wise anomaly abstention (PAL) loss ℓ_{pal} learns to abstain the prediction of outlier pixels from \mathbf{x}^{out} containing OE objects (i.e., cyan coloured masks) and calibrate the logit of inlier classes (i.e., reduction of the inlier logits) from both inlier image \mathbf{x}^{in} and outlier image \mathbf{x}^{out} . The EBM loss ℓ_{ebm} pushes the free energy E_θ to low values for inlier pixels and pulls that to high values for outlier pixels, where a regularisation loss ℓ_{reg} enforces the smoothness and sparsity constraints on the energy maps. Such EBM learning reduces the logit of inlier classes to share similar values at the same time, facilitating the ℓ_{pal} learning. Then, the pixel-wise penalty a_ω associated with the abstention class at position ω is estimated to bias the penalty to be low for outlier pixels and high for inlier pixels, which in turn encourages high free energy for anomalies and enforces ℓ_{pal} to abstain the anomalous pixels.	158
11.3	Confidence calibration performances between WideResnet34 baseline, Meta-OoD [27], and our approach.	167
A.1	Example of the multi-scale structural and non-structural anomaly localisation result for an MVTec AD [13] image, using both the local and global IGD models. The global model tends to produce smooth results but with some mistakes, while the local model produces jagged results, but without the global mistakes, so by combining the two results, we obtain a smooth and correct anomaly heatmap.	174
A.2	t-sne visualisation from MVTec (class bottle).	176

A.3	Qualitative visual results from Hyper-Kvasir testing set (red = anomaly).	179
A.4	Qualitative results of our anomaly localisation results on the MVTEC AD testing set (red = high probability of anomaly).	182
B.1	AUC w.r.t. top- k (Left) and the margin m (Right).	185
C.1	From the input image (anomaly highlighted with a yellow box), the initial prediction shows the original segmentation results with anoma- lies classified as a one of the pre-defined inlier classes. Anomaly pre- dictions from our method show an anomaly map with high scores (in yellow and red) for anomalous pixels. In our final prediction , anoma- lous pixels are coloured in cyan.	188

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Yu Tian

June 15, 2022

Acknowledgements

This thesis represents my work and milestones in more than three years of dedication at the University of Adelaide and specifically within the Australian Institute of Machine Learning (AIML). Since the first day of my Ph.D., I have felt home in Adelaide. I have been given unique and tremendous opportunities, guidance, encouragement from all the professors and colleagues at AIML.

First and foremost, I would thank my principal supervisor, Prof. Gustavo Carneiro, for his help during my Ph.D. study. He has been supportive since the days we met during my honors year. Ever since, Gustavo has supported me not only by providing mentorship over these years, but also be supported academically and emotionally through the rough road to finish this thesis. Without his help, I cannot succeed in my Ph.D. career. It's my great pride for being his student. To me, Gustavo is not only a supervisor, but also a role model in life, with his hard-working, kind, optimistic, and humble life attitudes.

I would also thank other members in my advisory team, Dr. Johan Verjans, Dr. Guansong Pang, and Prof. Rajvinder Singh, for their patient support and for all of the opportunities I was given to enhance my research. I would also like to thank them for sharing and deriving new ideas with me, for revising paper drafts, for explaining academic issues to me, for providing career advice, and for providing accompany during hard times.

I am also grateful to all my co-authors. I really miss the impressive discussion with them. These constructive discussions gave me much inspiration and resulted in many successful research projects and papers. They are Gabriel Maicas, Leonardo Zorron Cheng Tao Pu, Fengbei Liu, Yuyuan Liu, Yuanhong Chen, and Chong Wang. I want to thank other colleagues in my group and friends outside my group for chatting and sharing news about our life. They are always so supportive during my Ph.D. and I miss all those happy hours we spent together.

Last but not least, I would like to thank my family for their selfless love and support during my study. I can say I have a perfect family and they have given up many things for me to be at Adelaide to pursue my study abroad. They have cherished with me every great moment and supported me whenever I need it. Without them, I will not be the person I am today and I'm grateful for the comfortable and happy life they created to allow me to focus on my research.

Publications

This thesis contains the following works that have been published or prepared for publication (* indicates equal contribution):

- **Yu Tian**, Guansong Pang, Yuyuan Liu, Chong Wang, Yuanhong Chen, Fengbei Liu, Rajvinder Singh, Johan W Verjans, Gustavo Carneiro. Unsupervised Anomaly Detection in Medical Images with a Memory-augmented Multi-level Cross-attention Masked Autoencoder. *Arxiv Preprint, Under Review*, 2022.
- **Yu Tian***, Yuyuan Liu*, Guansong Pang, Fengbei Liu, Yuanhong Chen, Gustavo Carneiro. Pixel-wise Energy-biased Abstention Learning for Anomaly Segmentation on Complex Urban Driving Scenes. *Arxiv Preprint, Under Review*, 2022.
- **Yu Tian***, Fengbei Liu*, Guansong Pang, Yuanhong Chen, Yuyuan Liu, Johan W Verjans, Rajvinder Singh, Gustavo Carneiro. Self-supervised Pseudo Multi-class Pre-training for Unsupervised Anomaly Detection and Segmentation in Medical Images. *ArXiv Preprint, Under Review*, 2021.
- **Yu Tian**, Guansong Pang, Fengbei Liu, Yuyuan Liu, Chong Wang, Yuanhong Chen, Johan W Verjans, Gustavo Carneiro. Contrastive Transformer-based Multiple Instance Learning for Weakly Supervised Polyp Frame Detection. *International Conference on Medical Imaging Computing and Computer-Assisted Intervention (MICCAI)*, 2022, **Early Accept**.
- Yuanhong Chen*, **Yu Tian***, Guansong Pang, Gustavo Carneiro. Deep One-Class Classification via Interpolated Gaussian Descriptor. In *Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, 2022, **Oral**.
- **Yu Tian**, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, Gustavo Carneiro. Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning. In *International Conference on Computer Vision (ICCV)*, 2021.
- **Yu Tian**, Guansong Pang, Fengbei Liu, Yuanhong Chen, Seon Ho Shin, Johan W Verjans, Rajvinder Singh, Gustavo Carneiro. Constrained Contrastive Distribution

Learning for Unsupervised Anomaly Detection and Localisation in Medical Images. In *International Conference on Medical Imaging Computing and Computer-Assisted Intervention (MICCAI)*, 2021.

- **Yu Tian**, Gabriel Maicas, Leonardo Z.C.T. Pu, Rajvinder Singh, Johan W. Verjans, Gustavo Carneiro. Few-Shot Anomaly Detection for Polyp Frames from Colonoscopy. In *International Conference on Medical Imaging Computing and Computer-Assisted Intervention (MICCAI)*, 2020.
- Yuyuan Liu*, **Yu Tian***, Gabriel Maicas, Leonardo Z.C.T. Pu, Rajvinder Singh, Johan W Verjans, Gustavo Carneiro. Photoshopping Colonoscopy Video Frames. In *International Symposium on Biomedical Imaging (ISBI)*, 2020.

In addition, I have the following papers not included in this thesis:

- Fengbei Liu, Yuanhong Chen, **Yu Tian**, Yuyuan Liu, Chong Wang, Vasileios Belagiannis, Gustavo Carneiro. NVUM: Non-Volatile Unbiased Memory for Robust Medical Image Classification. *International Conference on Medical Imaging Computing and Computer-Assisted Intervention (MICCAI)*, 2022, **Early Accept**.
- Chong Wang, Yuanhong Chen, Yuyuan Liu, **Yu Tian**, Fengbei Liu, Davis McCarthy, Michael Elliott, Helen Frazer, Gustavo Carneiro. Knowledge Distillation to Ensemble Global and Interpretable Prototype-based Mammogram Classification Models. *International Conference on Medical Imaging Computing and Computer-Assisted Intervention (MICCAI)*, 2022, **Early Accept**.
- Yuanhong Chen, Wang Hu, Chong Wang, **Yu Tian**, Fengbei Liu, Yuyuan Liu, Michael Elliott, Davis McCarthy, Helen Frazer, Gustavo Carneiro. Multi-view Local Co-occurrence and Global Consistency Learning Improve Mammogram Classification Generalisation. *International Conference on Medical Imaging Computing and Computer-Assisted Intervention (MICCAI)*, 2022, **Early Accept**.
- Fengbei Liu*, **Yu Tian***, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, Gustavo Carneiro. ACPL: Anti-curriculum Pseudo-labelling for Semi-supervised Medical Image Classification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- Yuyuan Liu, **Yu Tian**, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, Gustavo Carneiro. Perturbed and Strict Mean Teachers for Semi-supervised Semantic Segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

- Fengbei Liu*, **Yu Tian***, Filipe R. Cordeiro, Vasileios Belagiannis, Ian Reid, Gustavo Carneiro. Self-supervised Mean Teacher for Semi-supervised Chest X-ray Classification. In *International Workshop on Machine Learning in Medical Imaging, MICCAI-MLMI*, 2021.
- Leonardo Z.C.T. Pu, Gabriel Maicas, **Yu Tian**, Takeshi Yamamura, Masanao Nakamura, Hiroto Suzuki, Gurfarmaan Singh, Khizar Rana, Yoshiki Hirooka, Alastair D. Burt, Mitsuhiro Fujishiro, Gustavo Carneiro, Rajvinder Singh. Computer-aided diagnosis for characterization of colorectal lesions: a comprehensive software including serrated lesions. In *Gastrointestinal Endoscopy (GIE)*, 2020.
- **Yu Tian**, Leonardo Z.C.T. Pu, Rajvinder Singh, Alastair D. Burt, Gustavo Carneiro. One-stage Five-class Polyp Detection and Classification. In *International Symposium on Biomedical Imaging (ISBI)*, 2019.
- Leonardo Z.C.T. Pu, Gabriel Maicas, **Yu Tian** and others. Prospective study assessing a comprehensive computer-aided diagnosis for characterization of colorectal lesions: Results from different centers and imaging technologies. In *Journal of Gastroenterology and Hepatology*, 2019.
- Yuanhong Chen, Fengbei Liu, **Yu Tian**, Yuyuan Liu, Gustavo Carneiro. Semantic-guided Image Virtual Attribute Learning for Noisy Multi-label Chest X-ray Classification. *Arxiv Preprint, Under Review*, 2022.
- Yuyuan Liu, **Yu Tian**, Chong Wang, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, Gustavo Carneiro. Translation Consistent Semi-supervised Segmentation for 3D Medical Images. *Arxiv Preprint, Under Review*, 2022.

Abstract

Anomaly detection is a fundamental problem in computer vision and medical imaging, which aims to detect unseen (i.e., not present in the training set) abnormal data instances that deviate from the distribution of seen (or present in the training set) normal instances. Deep neural networks have been the dominant model behind current solutions that have achieved great success in different application domains. Anomaly detection can be formulated as: (i) unsupervised anomaly detection (UAD) developed with a one-class classification method that only uses normal training data, (ii) few shot anomaly detection that uses a small amount of abnormal training data and a large amount of normal training data, and (iii) weakly supervised learning for video anomaly detection with video-level labels without any indication of where the anomaly happens inside the video sequence. Despite the remarkable achievements of current approaches, there are still many challenges worth exploring to advance the field.

Traditional reconstruction-based UAD methods use generative models to learn to reconstruct normal training images, where the assumption is that these models will reconstruct unseen abnormal images with larger error than the normal images. However, such an assumption often fails since modern generative models, such as autoencoders (AE) and generative adversarial networks (GAN), can generalise well to unseen abnormal images and yield low reconstruction errors, particularly for hard anomalies (i.e., subtle abnormal samples that look similar to normal instances). Thus, this thesis first targets this low reconstruction error for hard anomaly, present in generative models. We design several new reconstruction-based UAD methods that explicitly constrain the generative model to be able to only reconstruct normality patterns, reducing their ability to reconstruct unseen abnormal cases, and consequently improving their unsupervised anomaly detection accuracy. Moreover, we argue that another major issue that may reduce UAD accuracy is the inadequate feature representations obtained from pre-trained models designed to solve general classification tasks instead of UAD tasks. To address this issue, we propose the new self-supervised pre-training methods in the field designed specifically for downstream UAD tasks. When pre-training off-the-shelf anomaly classifiers, our self-supervised methods are shown to enable substantial improvements in terms of anomaly detection accuracy. We also notice that the accuracy of UAD methods can be improved by leveraging a few labelled abnormal samples during

training, which should be used in addition the normal samples to facilitate the classification of normal and abnormal instances. This idea allowed us to propose the new few-shot anomaly detection method to improve anomaly detection accuracy. Furthermore, we propose a new video anomaly detection approach that relies on weak video-level annotations. One of the major challenges of weakly supervised video anomaly detection (WVAD) is how to accurately identify anomalous frames or snippets from abnormal videos during training. Our solution for WAVD involves the design of a new temporal feature learning and a novel transformer-based multiple instance learning framework. Finally, we propose a simple and effective anomaly segmentation model that targets the pixel-wise anomaly detection task from complex urban driving scenes. This method aims to address the fundamental problem that current semantic segmentation models often produce misclassifications on unexpected road anomalies. We conduct our experiments on public anomaly detection and segmentation benchmarks and most of the methods presented in this thesis achieve state-of-the-art (SOTA) performance on various natural image and medical image analysis datasets.

Chapter 1

Introduction

Anomaly detection is a fundamental task in computer vision and medical image analysis, which consists of training a one-class classification model with a set of normal samples, and during testing this model must be able to contrast between normal and abnormal samples, even though the model has not been exposed to abnormal samples during training. The development of models and algorithms for anomaly detection has been an active research area, and current approaches play an important role on many real-world applications, such as automatic surveillance systems, medical malignant detection, safety for artificial intelligence systems, industrial defect detection, and etc. Traditional approaches formulate the problem as a one-class classification method using one-class SVM [35, 202] to learn a discriminative hyperplane to map the normal samples, or utilise clustering methods, such as k-means or Gaussian Mixture Models (GMM) [248, 272], to construct a normality distribution that identifies anomalies as samples that fall outside this distribution. However, these traditional methods show relatively poor performance when processing high-dimensional image and video data. With the development of deep learning methods, deep neural networks have become the main model explored for anomaly detection in both computer vision and medical image analysis, leading to superior performance in several benchmarks, compared with previously proposed machine learning models.

1.1 Anomaly Detection Setups

The task of anomaly detection has been intensively studied, where many problem setups have been proposed, which include: unsupervised anomaly detection (UAD), few-shot anomaly detection, and weakly-supervised anomaly detection (WAD), as shown in Fig. 1.1. Anomaly detection methods are usually applied to problems that are difficult to obtain high quality data and annotation. In computer vision, anomaly detection can be applied to detect abnormal events from video surveillance or detect industrial

defects from scanning images because the anomalies in these applications are often open-set, where it is hard to collect all possible categories for effectively identifying all unknown anomalies. Furthermore, in medical images, anomaly detection applied to disease screening problems is particularly interesting because most of the patients are healthy, where it is hard to collect a large number of abnormal/unhealthy data and annotations in real world clinical scenarios.

Unsupervised Anomaly Detection: UAD methods typically train a one-class classifier (OCC) using only data from the normal class, and anomalies (or abnormal cases) are detected based on the extent that testing samples deviate from the normal class. Such UAD formulation is crucial for detecting and segmenting anomalies in many applications (e.g., disease screening datasets in medical image analysis), which contain a disproportionately large number of normal (or healthy) images, and a small amount of abnormal (or disease) images. Not only is the collection and annotation of such heavily imbalanced training sets challenging, but it is also hard to acquire a representative dataset containing a reasonable number of images from all possible abnormal subclasses given the intrinsic variations in the visualisation of different anomalies [139, 218, 221]. This can also bring substantial benefits for computer vision applications, such as industrial defect detection [33] or detecting road anomalies for self-driving systems [49, 219], where those abnormalities are often rare to collect and annotate.

Few-shot Anomaly Detection: Unfortunately, in practice UAD methods can misclassify outliers that lie relatively close to inliers (e.g., when a lesioned tissue occupies a small area of the image). One possible way to address such problem is to leverage a small amount of abnormal training data through the design of training methods that can deal with heavily imbalanced learning problems [126, 128]. Even though they may be effective, these approaches still need a fairly high number of abnormal training images. Few-shot anomaly detection aims to propose a middle ground between these approaches to effectively address the issues of requiring a relatively large annotated data set from imbalance classification (containing normal and abnormal data), and misclassifying challenging outliers from UAD.

Weakly-supervised Anomaly Detection: Another major setup of anomaly detection is the weakly-supervised video anomaly detection (WVAD), where the training set contains video-level normal and abnormal labels, but no indication of where in the video the anomaly is present. Using this training data, the goal is to train a classifier that not only classifies a test video into normal or abnormal, but it also localises the anomaly within the video. WVAD has many applications, such as in the context of surveillance, where examples of anomaly are bullying, shoplifting, violence, etc., and in the context of colonoscopy videos, examples of anomaly are frames containing colon polyps. For both applications, it is of utmost importance to detect abnormal frames because surveillance and colonoscopy videos are often annotated with video-level labels in real-world datasets. Although aforementioned UAD setup trained exclusively with

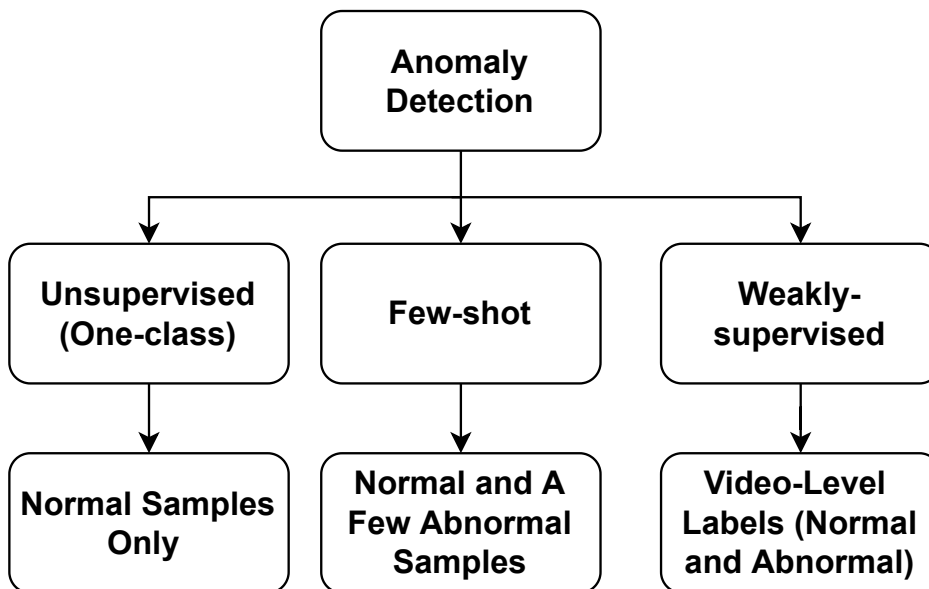


Figure 1.1: Taxonomy of three types of deep anomaly detection models explored in this thesis.

normal videos has been explored in this context [65, 87, 97, 145, 183, 184, 264], the best performing approaches explore a weakly-supervised setup using training samples with video-level label annotations [211, 217]. This weakly-supervised setup targets a better anomaly classification accuracy at the expense of a relatively small human annotation effort, compared with UAD approaches.

Even though previous approaches have shown accurate anomaly detection results, we still find significant challenges that need to be further studied from both empirical and theoretical viewpoints. In this thesis, we introduce our works under the aforementioned anomaly detection setups, i.e., unsupervised anomaly detection, few-shot anomaly detection and weakly supervised anomaly detection, to advance the field of detecting and localising abnormalities in industrial data, colonoscopy data, fundus data, Covid-19 Chest X-ray (CXR) data, self-driving obstacle data and surveillance data (See Fig. 1.2).

1.2 Motivation

Anomalies are by nature a rare event, which means that they are hard to find, particularly when compared with a massive amount of normal data. In addition, it will be even harder to collect a reasonable number of samples belonging to anomaly subclasses. Therefore, instead of trying to acquire a relatively balanced training set of

normal and abnormal samples, anomaly detection approaches either ignore abnormal data altogether or rely on extremely small amounts of (weakly-labelled) abnormal data. This effort reduction in terms of training data collection is the main motivation behind anomaly detection approaches.

Current UAD approaches [32, 34, 56, 139, 199, 216, 229] train deep generative models, such as autoencoder (AE) [108] and generative adversarial networks (GAN) [58], to reconstruct normal images, and anomalies are detected based on large reconstruction error values [165]. These approaches rely on a low-dimensional image representation that must be effective for reconstructing normal images, where the main challenge is on how to enforce this representation to not allow an accurate reconstruction of abnormal images, particularly the ones containing subtle anomalies. We argue that one way of improving the reconstruction-based UAD approaches is to use a dual GAN structure, which consists of two generators and two discriminators, to explicitly constrain both the latent and image spaces, yielding better anomaly detection performance (please see Chapter 3).

Inspired by recent developments in the field [10, 30, 77, 90, 94, 133], we propose the first¹ self-supervised learning for anomaly detection in medical images designed to tackle the low reconstruction error of subtle anomalies. This is achieved by pre-training the UAD models to learn fine-grained feature representations with the constrained contrastive distribution (CCD) model (please see Chapter 4). This method above is further extended to form a tighter and denser cluster than the CCD model in Chapter 5, where the cluster is formed by re-formulating the standard one-class UAD problem into an auxiliary multi-class centring/clustering problem. Another major issue of UAD approaches is that they often suffer from overfitting the training data, especially when the training set is small or contaminated with anomalous samples. This is of utmost importance in practice because real-world anomaly detection datasets often contain a small amount of anomalous contamination, challenging the effectiveness and robustness of existing UAD systems. To address this problem, we propose a new UAD model that learns a one-class Gaussian anomaly classifier trained with adversarially interpolated training samples to alleviate such issues, and for the first time, to assess the robustness of anomaly detectors to training sets that are small or contaminated with anomalous samples (please see Chapter 6). The subtle anomaly reconstruction error issue of UAD methods can also be addressed by training a transformer based memory-augmented masked autoencoder to explicitly encode and reconstruct normality patterns, thus forcing challenging anomaly cases to produce high reconstruction errors (please see Chapter 7).

Another way to improve the accuracy of UAD methods [56, 139, 148, 273], particularly with regards to subtle anomalies, is to include a small set of abnormal data into the training set. Such problem can be defined as a highly-imbalanced learning problem or a

¹To the best of our knowledge.

few-shot learning task. Imbalanced learning methods [126, 128] can in principle address this problem, but these approaches still need a much larger proportion and number of abnormal training images than is usually available in anomaly detection datasets. This thesis proposes the first middle-ground work between imbalanced learning and UAD. Our setup utilises a large quantity of normal data, and a comparatively much smaller amount of anomalous training samples, to effectively detect anomalies based on a novel few-shot anomaly detection method (Chapter 10). The resulting anomaly classifier requires significantly less abnormal training data to achieve better accuracy on anomaly classification than the traditional imbalance learning approaches, and at the same time it produces substantially more accurate results than UAD approaches.

Finally, this thesis explores weakly supervised video anomaly detection (WVAD) that is trained with normal and abnormal video-level labelled samples, and during testing, it aims to identify the time window when anomalous events happen in videos. One of the major challenges of WVAD is how to identify anomalous snippets from a whole video labelled as abnormal both during training and testing. This is due to the following two reasons: 1) the majority of snippets from an abnormal video contain normal events, which can overwhelm the training process and challenge the fitting of the few abnormal snippets; and 2) abnormal snippets may not be sufficiently different from normal ones, making a clear separation between normal and abnormal snippets challenging. To tackle such challenges, we propose a novel and theoretically sound method based on a feature magnitude learning function to recognise abnormal snippets, substantially improving the robustness to the normal snippets from abnormal videos (please see Chapter 8). We apply this WVAD method to polyp frame detection from weakly-labelled colonoscopy videos. Such setup consists of a vital clinical application for efficient and accurate colon cancer pre-diagnosis using minimally curated datasets directly available from hospitals and clinics (please see Chapter 9).

The thesis also investigates anomaly detection problems in self-driving systems, which produces pixel-wise anomaly classification for semantic segmentation models. Current segmentation models use common uncertainty measures (e.g., classification entropy or uncertainty) to detect anomalies [17, 20, 27, 49, 103, 130, 155, 246], but they often fail to properly recognise anomalous objects that deviate from the training inlier distribution (e.g., Cityscapes [39]), leading to potentially fatal model decisions. We take an alternative approach and propose a simple and effective anomaly segmentation method to tackle this task. Our method introduces a novel pixel-wise abstention learning to improve the precision and robustness to detect small anomalous objects, and achieve substantial performance improvements on existing benchmarks (please see Chapter 11).

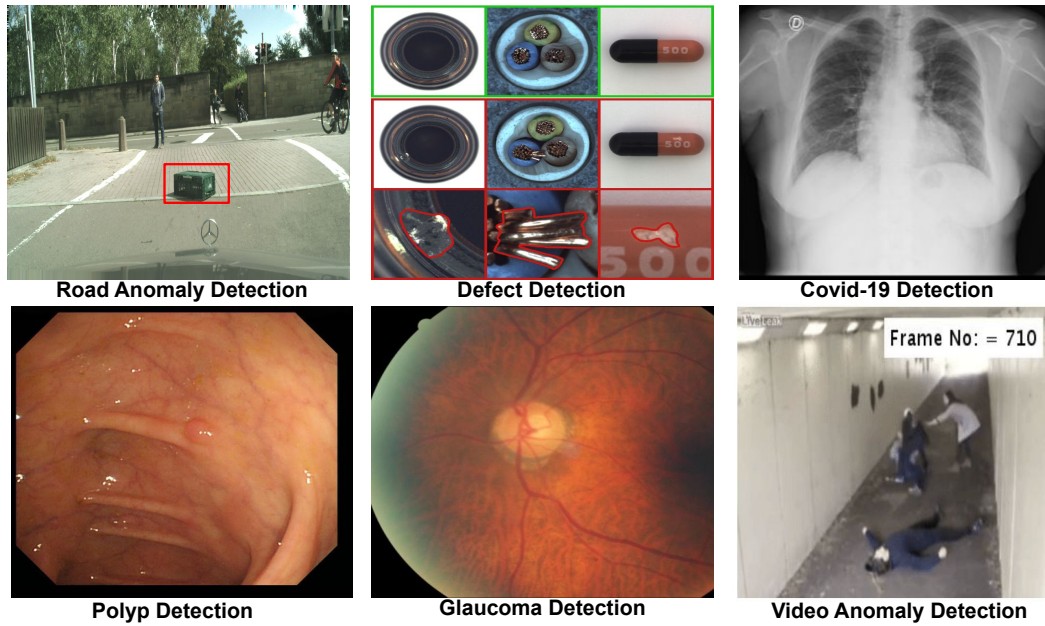


Figure 1.2: Anomaly detection applications explored in this thesis: road obstacles, industrial defect detection, Covid-19 detection from Chest X-ray, polyp detection from colonoscopy frames, glaucoma detection from fundus screening images, and violence detection from surveillance videos.

1.3 Contributions and Thesis Outline

We propose different deep learning methods for various anomaly detection tasks with unlabelled, weakly labelled and few-shot labelled data. This thesis aims to improve the performance (e.g. accuracy, robustness, stability and efficiency) of previous anomaly detection methods from the literature, and to propose new methodological anomaly detection formulations and novel computer vision and medical image analysis tasks.

The contributions of this thesis can be outlined as:

- **Chapter 2** provides the details about the related literature. We introduce previously published unsupervised anomaly detection methods and we also review few-shot and weakly supervised anomaly detection approaches. Furthermore, we present anomaly detection for semantic segmentation from complex urban driving scenes.
- **Chapter 3** describes our proposed anomaly detection generative adversarial network (ADGAN) approach for detecting anomalies from colonoscopy frames. Our ADGAN comprises two generators and two discriminators, which are designed to explicitly constrain the latent space, where the image GAN aims to preserve both the global

Anomaly Detection Setups	Computer Vision	Medical Imaging
Unsupervised Anomaly Detection	Chapter 6	Chapter 3, 4, 5, 6, 7
Weakly Supervised Anomaly Detection	Chapter 8	Chapter 9
Few-shot Anomaly Detection	Chapter 10	-
Pixel-wise Anomaly Detection in Self-driving	Chapter 11	-

Table 1.1: **Thesis contributions:** We propose different deep learning methods for various anomaly detection tasks, including unsupervised anomaly detection without abnormal training data, weakly supervised anomaly detection with weakly labelled video training data, and few-shot anomaly detection with few labelled image data.

and local information of input image data by combining mean square error (MSE) and binary cross entropy (BCE) losses for better reconstruction of normal images. We show that our ADGAN is more effective and more accurate than previous state-of-the-art (SOTA) methods.

- **Chapter 4** shows that one of the major challenges that hinders the accuracy of UAD methods is the difficulty of learning effective low-dimensional image representations to detect and segment subtle anomalies. To address this issue, we propose the first self-supervised pre-training method specifically designed for UAD in medical imaging, named constrained contrastive distribution (CCD), which learns fine-grained feature representations by simultaneously predicting the distribution of augmented data and image contexts using contrastive learning with pretext constraints. We show that the pre-trained model can be adapted to a wide variety of anomaly classifiers, yielding better improvements than with Imagenet pre-trained models.
- **Chapter 5** proposes a novel self-supervised pre-training method specifically designed for MIA UAD applications to form denser and tighter clusters for normal sample representations. Our method introduces a new design of the contrastive learning optimisation that converts the OCC problem into a multi-class clustering task with the help of our MedMix augmentations that simulate different types of lesions of varying size and appearance using only normal training data. The proposed approach is shown to learn effective feature representations that can adapt well to different types of downstream UAD tasks and is able to be applied to several MIA problems.
- **Chapter 6** focuses on building a new UAD approach for image anomaly detection. The new one-class classifier (OCC) model targets the learning of an effective normality descriptor with a theoretically sound derivation of the expectation-maximisation (EM) algorithm that optimises a Gaussian anomaly classifier constrained by adversarial interpolation and multi-scale image reconstruction. This results in a robust anomaly classifier that can be trained with small and contaminated training data.

Our work introduces a new benchmark for UAD approaches with training sets that are small or contaminated with anomalous samples.

- **Chapter 7** investigates, for the first time in the field, the potential effectiveness of masked autoencoders (MAE) for anomaly detection. Reconstruction methods, which detect anomalies from image reconstruction errors, are advantageous for medical imaging tasks but often fail because they can have low reconstruction errors even for anomalous images. In this chapter, we propose a new reconstruction-based UAD approach that addresses this low-reconstruction error issue for anomalous images by using transformers for the encoder and decoder architectures. We further introduce a novel memory module to facilitate the MAE training. The method achieves SOTA performance on two MIA applications.
- **Chapter 8** proposes a weakly-supervised anomaly detection method that can identify anomalous snippets from a whole video labelled as abnormal both during training and testing. To accurately identify those anomalous snippets, we propose the robust temporal feature magnitude (RTFM), which substantially improves the robustness of current multiple instance learning (MIL) approaches trained with video-level weak labels. A new multi-scale temporal feature learning is also introduced to seamlessly incorporate long and short-range temporal dependencies within each video. We also show a detailed theoretical analysis of our RTFM algorithm. The resulting model is shown to achieve SOTA accuracy on four video surveillance datasets.
- **Chapter 9** argues that current polyp detection methods from colonoscopy videos often ignore the importance of temporal information in consecutive video frames. Hence, in this chapter, we formulate polyp detection as a weakly-supervised anomaly detection task that uses video-level labelled training data to detect frame-level polyps. This is the first work to consider polyp detection as weakly-supervised anomaly detection task and we introduce a novel contrastive snippet mining to enable the detection of challenging polyp cases (e.g., small, flat, or partially visible polyps). This method is validated on a new large scale colonoscopy video dataset and achieves the best results when compared with previous leading approaches.
- **Chapter 10** introduces a new few-shot anomaly detection network (FSAD-Net) based on an encoder trained to maximise the mutual information between feature embeddings and normal images, followed by a few-shot score inference network, trained with a large set of normal samples and a substantially smaller set of abnormal samples. Our model is designed to improve the accuracy of UAD approaches without requiring a relatively large amount of labelled abnormal training samples, as needed by imbalanced learning approaches.
- **Chapter 11** presents a fundamental issue with current semantic segmentation methods that tend to produce inaccurate predictions of unexpected road anomalous ob-

jects (e.g., unexpected objects in the middle of the road can be mis-classified as the road class). We propose a new method, named pixel-wise energy-biased abstention learning (PEBAL), to address such failure cases for existing semantic segmentation models. The proposed PEBAL explores a nontrivial joint training between a novel pixel-wise abstention learning (PAL) that learns an adaptive pixel-level anomaly class, and an energy-based model (EBM) that learns inlier pixel distribution. PEBAL achieves substantial performance improvement for detecting anomalous objects from real-world urban driving scenes compared with previous SOTA.

- **Chapter 12** summarises the methods and the contributions of this thesis and discusses the potential future directions for anomaly detection.

Chapter 2

Literature Review

2.1 Unsupervised Anomaly Detection

Anomalies are defined as data that do not conform to the general distribution of normal data [166]. For instance, in medical image analysis, anomaly detection can be used in the detection of abnormal lesions in normal tissue [174, 199]. Unsupervised anomaly detection (UAD) is the dominant method to tackle such problem, which are generally formulated with a one-class classifier [23] that is trained using only normal training samples. Important examples of UAD methods are provided by You et al. [253], who train an outlier detection method that combined sample representations and random walks on a representation graph. Also, Sabokrou et al. [195] propose a framework that consists of two models, where one model works as the novelty detector and the other supports it by improving the separability between enhanced normal samples and distorted anomalies.

Alternatively, some recent approaches use features extracted from pre-trained deep neural networks [72, 101, 205, 265], and train an anomaly score classifier using the extracted features. DSVDD [191] consider UAD as a one-class classification problem, which forces normal image features to be inside a hyper-sphere with a pre-defined centre and a radius that is minimised to include all training images. Markovitz et al. [150] extract human pose graphs from surveillance videos and cluster them in a latent space to distinguish between normal and abnormal. Morais et al. [153] propose an anomaly classifier to detect human-based anomalies using skeleton trajectories. The anomaly detection approaches based on ImageNet pre-trained models [42, 185] often fail to transfer representations learned from natural images to medical images.

UAD accuracy has recently been improved with self-supervised pre-training that relies on pretext tasks, consisting of predicting geometric or brightness transformations [10, 31, 77, 90, 94, 133] to learn fine-grained normality training features. For example, Pang et al. [169] propose the use of self-supervised learning to assign pseudo

labels based on the predicted anomaly scores. However, self-supervised learning methods [31, 90, 133] are designed to work for downstream multi-class classification problems, so there are no guarantees that such approaches will seamlessly translate to downstream UAD problems [238].

Recent UAD approaches also rely on generative models (i.e., GAN, autoencoder) for accurate anomaly detection, where the generated normal image is produced conditioned on another normal image. Autoencoder and GAN strategies assume that the abnormal data cannot be reconstructed correctly during testing given that the model is trained only with normal data. Liu et al. [65] propose a method based on future frame prediction in a video sequence using a GAN-based method trained with normal data only, and tested to distinguish normal and abnormal events on surveillance dataset. This approach is not applicable to medical images, and in particular colonoscopy data, because the scopes used to acquire such images can have quick and unpredictable motions, so future frames tend to be not as predictable as in a natural image sequence. Gong et al. [56] propose a memory-module to enforce the autoencoder to reconstruct samples into its normal form. Park et al. [170] improve such memory block by introducing a more complex and effective memory update mechanism and two new losses to optimise the memory module. OCGAN [175] introduce a few techniques to constrain the latent space to represent the normal class. Abati et al. [2] design an autoencoder based anomaly detection with a parametric density estimator that learns the probability distribution underlying its latent representations through an autoregressive procedure. Another GAN-based anomaly detection, Anogan [64], trains a DCGAN [55] network on normal (healthy) retinal OCT images. During testing, a computationally expensive iterative back-propagation process is run to produce the closest image to the input image. Schlegl et al. [199] addressed this large computational run-time issue by training an encoder after training a WGAN [5] to speed up the inference time. Replacing the iterative back-propagation from Anogan to an efficient encoding mechanism reduces inference running time, but introduces an ineffective two-stage training process.

Given that anomalies tend to be localised in a small region of the image that can be otherwise considered to be normal, it is important to pay attention not only to the detection, but also the localisation of anomalies. Unsupervised anomaly localisation targets the segmentation of anomalous pixels or patches, containing, for example, lesions in medical images [122], defects in industry images [11, 13], or road anomalies in traffic images [173, 219]. The main idea explored in anomaly localisation is based on extending the image based OCC/UAD to a pixel-based OCC/UAD, where testing produces a pixel-wise anomaly score map [8, 15]. In general, methods that can localise anomalies [11, 228] are tuned to particular range of anomaly sizes and structure, which can cause then to miss anomalies outside that range. Therefore, it is important to study new approaches that can localise anomalies of several sizes.

2.2 Weakly Supervised and Few-shot Anomaly Detection

Leveraging a few labelled abnormal samples during training has shown to provide substantial improvements over the aforementioned UAD approaches [137, 164, 193, 211, 216, 245, 257, 258, 259]. For example, Ruff et al. [193] propose to use a small pool of labelled abnormal samples to learn an end-to-end deep anomaly classifier for images. However, this method still requires a relatively large amount of abnormal samples. Hence, an important open research question is how to enable the training of anomaly detection models using significantly less anomalous images than imbalanced learning and previous anomaly detection approaches. Such strategy is named few-shot anomaly detection, which relies on a handful of anomalous images for training.

For video anomaly detection, it is too expensive to acquire large-scale frame-level label annotation. Hence, current SOTA video anomaly detection approaches rely on weakly supervised training that uses cheaper video-level annotations. Sultani et al. [211] proposed the use of video-level labels and introduced the large-scale weakly-supervised video anomaly detection dataset, UCF-Crime. Since then, this research direction has attracted growing attention from the community [73, 231, 245, 262]. Weakly-supervised video anomaly detection (WVAD) methods are mainly based on the multiple instance learning (MIL) framework [211]. However, most MIL-based methods [211, 262, 270] do not explicitly address the problem of label noise that might be present in a positive bag because a normal snippet can be mistakenly selected as the top abnormal event in an anomaly video. To deal with this problem, Zhong et al. [266] reformulated this problem as a binary noisy-label classification and used a graph convolution neural (GCN) network to fix the label noise. Although this paper shows more accurate results than [211], the training of GCN and MIL is computationally costly, and it can lead to an unconstrained latent space (i.e., normal and abnormal features can lie at any place of the feature space) that can cause unstable performance. The most recent work in this area [73] proposes a MIL self-training framework to refine snippet-level feature representations between weak labelled normal and abnormal videos.

However, those MIL based WVAD approaches suffer from some common issues, namely: 1) the top anomaly score in an abnormal video may not be from an abnormal snippet; 2) normal snippets randomly selected from normal videos may be too easy to fit, which challenges training convergence; 3) if the video has more than one abnormal snippet, we miss the chance of having a more effective training process containing more abnormal snippets per video; 4) the use of classification score provides a weak training signal that does not necessarily enable a good separation between normal and abnormal snippet; and 5) the identification of challenging abnormal snippets that have subtle anomalies is difficult and often incorrect. These issues above point to interesting

research problems that need further investigation.

2.3 Pixel-wise Anomaly Detection in Semantic Segmentation

Recent advances in semantic segmentation have shown tremendous improvements on complex urban driving scenes [116]. Despite the accurate predictions on the inlier classes, the model fails to properly recognise anomalous objects that deviate from the training inlier distribution such as the unexpected objects in the middle of the road. Current methods can be categorised to either uncertainty based or reconstruction based approaches. Early uncertainty-based methods [95, 118, 127] focused on the estimation of image-level anomalies and naively adapt their approach to pixel-wise task, which tended to mis-classify some hard pixels (e.g., inlier object boundaries and subtle anomalies) into inlier classes [103]. Jung et al. [103] mitigate such issue by iteratively replacing false anomalous boundary pixels with neighbouring non-boundary pixels that have low anomaly score, allowing the detection of anomalies without model re-training or adding extra models. Moreover, the aforementioned boundary issue is also alleviated with a pixel-wise uncertainty estimated with MC dropout in [106, 115, 155], but they still yield a poor pixel-wise anomaly detection accuracy [130]. Without fine-tuning using a proxy outlier dataset, the uncertainty estimation may not be accurate enough to detect anomalies, especially for the challenging abnormal cases that share similar appearance features with the normal objects.

Another strategy for pixel-wise anomaly detection is based on the use of extra models for image reconstruction. With such models, unseen abnormalities can be segmented from the reconstruction errors between the input image and its re-synthesised version based on its predicted segmentation map [8, 33, 40, 49, 86, 130, 230, 246]. Those approaches are challenged by the dependence on an accurate segmentation prediction, by the complexity of reconstruction models that usually require long training and inference processes, and also by the low quality of the reconstructed images. Moreover, reconstruction methods that rely on a discrepancy module require re-training whenever the inlier segmentation model changes due to input distribution shift [49], limiting their applicability in real-world systems.

Therefore, inspired by the image-level out-of-distribution detection [96], some pixel-wise anomaly detectors utilise the outlier exposure (OE) strategy that uses an auxiliary dataset of outliers, which has no overlap with real outliers (i.e., anomalies), to improve the anomaly detection performance. The leading approaches in this field adopt ImageNet [17, 18, 226], void class of Cityscape [49] or COCO [27] as the OE samples/pixels, where the expectation is that the model trained with the OE strategy can generalise to unseen outliers. However, maximising uncertainty for outliers using the OE strategy

may often lead to a deterioration of the inlier segmentation accuracy [18, 226]. Another major issue that affects the methods above is that the training set of OE samples often contains a disproportionately high amount of outliers [27], which can bias the segmentation toward the anomaly class, leading to poor anomaly segmentation accuracy. Therefore, even though the OE strategy has led to the development of successful methods, the issues around uncertainty maximisation and the large amount of outliers should be addressed to enable further improvements on pixel-wise anomaly detection.

2.4 Anomaly Detection Datasets

In this thesis, we conduct experiments using several publicly available datasets. For UAD datasets in computer vision, MNIST [48], Fashion MNIST [247] and CIFAR10 [113] have been widely used to benchmark image anomaly detection methods, and we follow the same experimental protocol as described in [1, 11, 41, 56, 176, 191, 228]. CIFAR10 contains 60,000 images with 10 classes. MNIST and Fashion MNIST contains 70,000 images with 10 classes of handwritten digits and fashion products, respectively. For each dataset, the benchmark consists of 10 anomaly detection experiments by considering images from each class as normal samples and the images from remaining classes as anomalies. The sampled normal data is split into training and testing sets with a ratio of 2:1. The test sets contain 10,000 abnormal samples from the remaining anomalous classes. The results reported in this thesis are the mean over the 10 anomaly detection experiments.

MVTec AD [13] is a recently released dataset that contains 5,354 high-resolution real-world images of 15 different industry objects and textures. The normal class of MVTEC AD is formed by the images without defects and consists of 3,629 images for training and 467 images for testing. The anomalous class has more than 70 categories of defects (such as dents, structural fails, contamination, etc.) and contains 1,258 images used only for testing. Furthermore, MVTEC AD also provides pixel-wise ground truth annotations for all anomalies, allowing the evaluation of not only anomaly detection, but also anomaly localisation.

We use four medical UAD datasets, namely: the colonoscopy images of Hyper-Kvasir dataset [21], Liu et al.’s colonoscopy dataset [139], the LAG glaucoma dataset using fundus images [121], and Covid-19 chest X-ray dataset [235]. Hyper-Kvasir is a large multi-class public gastrointestinal image dataset [21]. The data were collected from the gastroscopy and colonoscopy procedures from Baerum Hospital in Norway. All labels were produced by experienced clinicians. The dataset contains 110,079 images from abnormal (i.e., unhealthy) and normal (i.e., healthy) patients, where 10,662 of those images have been labelled, and each image has size 300×300 pixels. We use a subset of the normal (i.e., healthy) images from the dataset for training. Specifically, 2,100 images from ‘cecum’, ‘ileum’ and ‘bbps-2-3’ are selected as normal, from which

we use 1,600 for training and 500 for testing. We also take 1,000 abnormal images and their segmentation masks of polyps to be used exclusively for testing. LAG is a large scale fundus image dataset for glaucoma diagnosis [121], containing 4,854 fundus images with 1,711 positive glaucoma scans and 3,143 negative glaucoma scans, where images have size of 500×500 pixels. For the experiments, we use 2,343 normal (i.e., negative glaucoma) images for training, and 800 normal images and 1,711 abnormal images with positive glaucoma diagnosis with attention maps annotated by ophthalmologists. The attention maps are built using an eye tracking device, which automatically outputs a region of interest for glaucoma diagnosis [121]. Covid-X [235] has a training set with 1,670 COVID-19 positive chest X-ray images, and 13,794 COVID-19 negative chest X-ray images of size 299×299 pixels. The test set contains 400 chest X-rays, consisting of 200 positive and 200 negative images. We train the methods with the 13,794 COVID-19 negative chest X-ray training images and test on the 400 chest X-ray images. Liu et al.’s colonoscopy dataset is a colonoscopy image dataset with 18 colonoscopy videos from 15 patients [139]. The training set contains 13,250 normal (healthy) images without any polyps, and the testing set contains 967 images, with 290 abnormal images with polyps and 677 normal (healthy) images without polyps, where images have size 64×64 pixels. For few-shot anomaly detection, we modify Liu et al.’s colonoscopy dataset, and build a training set with 13,250 normal images (without polyps) and 10 to 80 abnormal images. The testing set contains 967 images, with 217 (25% of the set) abnormal images and 700 (75% of the set) normal images.

For the evaluation of weakly supervised anomaly detection approaches, four computer vision datasets are used, namely: ShanghaiTech [65], UCF-Crime [211], XD-Violence [245] and UCSD-Peds [250]. ShanghaiTech is a medium-scale dataset obtained from fixed-angle street video surveillance. It has 13 different background scenes and 437 videos, including 307 normal videos and 130 abnormal videos. The original dataset [65] is a popular benchmark for the anomaly detection task. Zhong et al. [266] reorganised the dataset by selecting a subset of the abnormal testing videos to insert into the training data to build a weakly supervised training set, so that both training and testing sets cover all 13 background scenes. We use the same procedure as in [266] to convert ShanghaiTech to the weakly supervised setting. UCF-Crime is a large-scale anomaly detection dataset [211] that contains 1,900 untrimmed videos with a total duration of 128 hours from real-world street and indoor surveillance cameras. Unlike the static backgrounds in ShanghaiTech, UCF-Crime consists of complicated and diverse backgrounds. Both training and testing sets contain the same number of normal and abnormal videos. The dataset covers 13 classes of anomalies in 1,610 training videos with video-level labels and 290 test videos with frame-level labels. XD-Violence is a recently proposed large-scale multi-scene anomaly detection dataset, collected from real-life movies, online videos, sport streaming, surveillance cameras and CCTVs [245]. The total duration of this dataset is over 217 hours, containing 4,754 untrimmed videos

with video-level labels in the training set and frame-level labels in the testing set. It is currently the largest publicly available video anomaly detection dataset. UCSD-Peds is a small-scale dataset comprising two sub-datasets: Ped1 with 70 videos and Peds2 with 28 videos. Previous papers [88, 266] re-formulate the dataset to the weakly supervised anomaly detection problem by randomly selecting 6 abnormal videos and 4 normal videos to the training set, and the remaining videos to the testing set. Results are computed with the mean accuracy over 10 times of this process.

For WVAD tasks in medical images, we propose a real-world large-scale video polyp detection dataset, containing colonoscopy videos collected from two widely used public datasets: Hyper-Kvasir [21] and LDPolypVideo [147]. The new dataset contains 61 normal videos without polyps and 102 abnormal videos with polyps for training, and 30 normal videos and 60 abnormal videos for testing. The videos in the training set have video-level labels and the videos in testing set contain frame-level labels. This dataset contains over one million frames and has diverse polyps with various sizes and shapes, making it one of the largest and most challenging colonoscopy datasets in the field.

Statement of Authorship

Title of Paper	Photoshopping Colonoscopy Video Frames
Publication Status	<input checked="" type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	Published in International Symposium on Biomedical Imaging (ISBI) 2020.

Principal Author

Name of Principal Author (Candidate)	Yu Tian			
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper.			
Overall percentage (%)	60			
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.			
Signature	<table border="1"> <tr> <td>_____</td> <td>Date</td> <td>09/03/2022</td> </tr> </table>	_____	Date	09/03/2022
_____	Date	09/03/2022		

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Yuyuan Liu			
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper.			
Signature	<table border="1"> <tr> <td>_____</td> <td>Date</td> <td>09/03/2022</td> </tr> </table>	_____	Date	09/03/2022
_____	Date	09/03/2022		

Name of Co-Author	Gabriel Maicas			
Contribution to the Paper	Discussion and writing the revision.			
Signature	<table border="1"> <tr> <td>_____</td> <td>Date</td> <td>09/03/2022</td> </tr> </table>	_____	Date	09/03/2022
_____	Date	09/03/2022		

Please cut and paste additional co-author panels here

Name of Co-Author	Leonardo Z.C.T. Pu		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Rajvinder Singh		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Johan W Verjans		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Please cut and paste additional co-author panels here as required.

Chapter 3

Photoshopping Colonoscopy Frames

Abstract

The automatic detection of frames containing polyps from a colonoscopy video sequence is an important first step for a fully automated colonoscopy analysis tool. Typically, such detection system is built using a large annotated data set of frames with and without polyps, which is expensive to be obtained. In this paper, we introduce a new system that detects frames containing polyps as anomalies from a distribution of frames from exams that do not contain any polyps. The system is trained using a one-class training set consisting of colonoscopy frames without polyps – such training set is considerably less expensive to obtain, compared to the 2-class data set mentioned above. During inference, the system is only able to reconstruct frames without polyps, and when it tries to reconstruct a frame with polyp, it automatically removes (i.e., photoshop) it from the frame – the difference between the input and reconstructed frames is used to detect frames with polyps. We name our proposed model as anomaly detection generative adversarial network (ADGAN), comprising a dual GAN with two generators and two discriminators. To test our framework, we use a new colonoscopy data set with 14317 images, split as a training set with 13350 images without polyps, and a testing set with 290 abnormal images containing polyps and 677 normal images without polyps. We show that our proposed approach achieves the state-of-the-art result on this data set, compared with recently proposed anomaly detection systems.

3.1 Introduction

Colorectal cancer is considered to be one of the most harmful cancers – current research suggests that it is the third largest cause of cancer deaths [62, 67]. Early detection of colorectal cancer can be performed with the colonoscopy procedure for at-risk patients

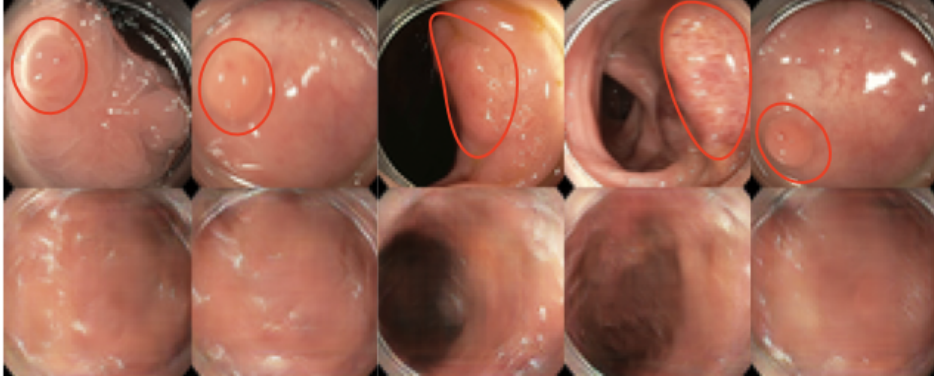


Figure 3.1: Top row shows test images containing polyps (highlighted with a red ellipse), which are considered to be anomalies in our framework. Bottom row shows the reconstructed images by our ADGAN model, which deviate with their top row input images leading to high reconstruction errors. Note that given that the ADGAN model was trained with images without polyps, it is biased to reconstruct images without polyps, as clearly seen in these examples.

with symptoms like hematochezia and anemia [59]. Colonoscopy is based on the navigation of a small camera in the colon that enables doctors to classify and possibly remove or sample polyps, which are considered as the precursors of colon cancer [62]. The accurate detection of colon polyps may improve 5-year survival rate to over 90% [62]. Unfortunately, the accuracy of such manual detection varies substantially, leading to potentially missing detection that can have harmful consequences for the patient [61]. For instance, the false negative polyp detection can lead to a future colon cancer, which can be dangerous or even fatal [225]. Therefore, automated detection of polyps is important in assisting doctors during a colonoscopy exam.

The automated polyp detection starts with the identification of frames containing polyps. Typically, such detection system consists of a 2-class classifier, trained with images containing polyps and images that do not contain polyps. The acquisition of such training set is expensive, requiring the manual annotation of large amounts of images for both classes. Furthermore, given the intrinsic variations in the visualisation of polyps, it is challenging to collect a training set that is rich enough to thoroughly represent the class of images that contain polyps [67]. To solve these two issues, this detection problem can be re-formulated as an anomaly detection problem, generally designed as a one-class classification problem that relies on a training set containing images that do not show the anomaly to be detected (i.e., the negative images) [23]. Such classification approach generally does not scale well with the size of the training set, reducing its applicability to large-scale medical image analysis problems. An alternative approach that addresses this scalability issue is based on a reconstruction model

that first trains an encoder-decoder with negative images. Such model will produce high reconstruction errors for positive images (i.e., images containing polyps) during testing stage since only negative images were used during training [56, 148, 273] – see Fig. 11.1. Note that the approaches above were developed for non-medical image analysis problems. These encoder/decoder approaches suffer from two issues: 1) the reliance on mean square error (MSE) loss to compute the distance between the reconstructed images and its original image, can only preserve local visual information [58, 148]; and 2) the latent space learned with encoder/decoder approaches can accurately reconstruct abnormal images with similar appearance features to normal images, leading to relatively small deviation between the distributions of normal and abnormal data. For instance, the model that learns with colon wall images can also reconstruct the abnormal colon wall image containing small polyps [174].

Anomaly detection can also be based on generative adversarial network (GAN) [58], which addresses the local visual information issue mentioned above with an adversarial training of a generator that tries to fool a discriminator to be confused with the classification between real and synthetic images. However, unlike the encoder-decoder model, GAN [58] cannot reconstruct an image based on a compressed latent variable from a given input image, and suffers from unstable training. Schlegl et al. [199] train an encoder to directly map an input image to GAN’s latent space and tackle the unstable training issue by replacing the vanilla GAN [58] with Wasserstein GAN (WGAN) [5]. Nevertheless, the framework proposed in [199] adopts a two-stage training strategy (encoder and WGAN are trained separately). Furthermore, the issues of the encoder/decoder model mentioned above are not addressed in [199].

In this paper, we propose a new WGAN-based [5] anomaly detection model that comprises two generators and two discriminators – this model is named anomaly detection GAN (ADGAN). Comparing with its competing approaches, our proposed model can produce an explicitly constrained latent space using latent generator and discriminator, and take advantage of GAN’s generation ability to preserve both global and local information of input data by combining MSE and binary cross entropy (BCE) losses to improve the performance of anomaly detection. We show that our ADGAN is more effective (by relying on a one-stage end-to-end training) and more accurate (for anomaly classification) than previous methods [174, 199]. We demonstrate that our method can reconstruct the abnormal images with polyps into normal images by automatically removing (i.e., ‘photoshopping’) the polyps (Fig. 11.1). These results are demonstrated on a new colonoscopy data set, containing 14317 high-quality colonoscopy images. The training set contains 13250 normal (healthy) images without polyps and we use 100 normal images as validation set. The testing set contains 290 abnormal images with polyps and 677 normal images without polyps (i.e., 30% of the testing images are abnormal).

3.2 Related work

Anomalies are defined as data that do not conform to the general distribution of normal data. In medical image analysis, anomaly detection can be used, for example, in the detection of abnormal lesions in normal tissue [174, 199]. Anomaly detection models are generally based on one-class classifiers [23], encoder-decoder models [60, 174] and GAN approaches [174, 199]. One-class classifiers are generally based on Gaussian processes, which do not scale well with training set size [23]. Encoder/decoder and GAN strategies assume that the abnormal data cannot be reconstructed correctly during testing stage given a model trained only with normal data. The typical encoder-decoder model for anomaly detection learns a deep auto-encoder [60] from the normal data during training stage, and during testing, this model is expected to produce larger reconstruction error for abnormal inputs [68], hopefully containing the lesions of interest. The main issue with the encoder-decoder method is that the trained model tends to accurately reconstruct abnormal samples during testing, leading to relatively small deviation between the distribution of normal and abnormal images.

GAN-based models usually involve a conditional GAN approach, where the generated normal image is produced conditioned on another normal image. For instance, Liu et al. [65] proposed a method based on future frame prediction in a video sequence using a GAN-based framework trained with normal data only, and tested to distinguish normal and abnormal events on surveillance data set. This approach is not applicable to medical image, and in particular colonoscopy data, because future frames tend to be not as predictable from past frames in a sequence. OCGAN [174] was proposed to distinguish abnormal data using a framework comprised of a denoising auto-encoder network, latent discriminator, visual discriminator and a classifier. Nonetheless, the experiment results of this work indicate that the model perform unsatisfactorily on complex image data (e.g., surveillance and medical images) [174]. Another GAN-based anomaly detection, Anogan [64], trains a DCGAN [55] network on normal (healthy) retinal OCT images. During testing, a computationally expensive iterative back-propagation process is run to produce the closest image to the input image. Schlegl et al. [199] addressed this large computational run-time issue by training an encoder after training a WGAN [5] to speed up the inference time. Replacing the iterative back-propagation from Anogan to an efficient encoding mechanism reduces inference running time, but introduces an ineffective two-stage training process. Another problem with the training above is that the encoder is under-constrained given that the MSE loss only recovers local visual information and misses global information.

By taking the motivation from [199], our proposed GAN framework, based on the Wasserstein GAN (WGAN) [5], resolves the issues mentioned above by an end-to-end (i.e., one-step) training of a dual GAN that uses a new loss function that minimises both the local and global reconstruction errors.

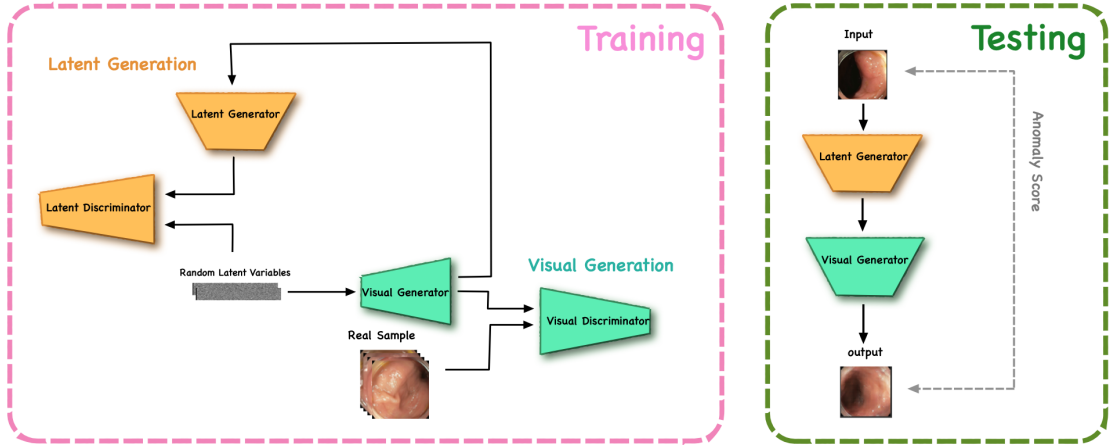


Figure 3.2: Our proposed ADGAN model trains the visual generator, visual discriminator, latent generator and latent discriminator using adversarial training (left). During testing, the input image is processed by the latent generator and the produced latent embedding is used by the visual generator to produce the output image, which is then compared with the input image to compute the anomaly score.

3.3 Data Set and Methods

3.3.1 Data Set

The data set is obtained from 18 colonoscopy videos from 15 patients. Video frames containing blurred visual information are removed using the variance of Laplacian method [66]. We then sub-sample consecutive frames by taking one frame every five frames because consecutive frames generally contain similar visual information that makes GAN training ineffective. We also remove frames containing feces and water to improve the training efficiency (we plan to deal with such distractors in future work). As a result, the frames used for training and testing are sharp, clean (of feces and water), and discontinuous (in time domain).

This data set is defined by $\mathcal{D} = \{\mathbf{x}_i, d_i, y_i\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x} : \Omega \rightarrow \mathbb{R}^3$ denotes a colonoscopy frame (Ω represents the frame lattice), $d_i \in \mathbb{N}$ represents patient identification¹, $y_i \in \mathcal{Y} = \{Normal, Abnormal\}$ denotes the normal healthy colorectal frames and abnormal colorectal frames that contains polyp. The distribution of this data set is as follows: 1) Training set: 13250 normal (healthy) images without any polyps; 2) Validation set: 100 normal (healthy) images for model selection; and 3) Testing set: 967 images, with 290 (30% of the set) abnormal images with polyps and 677 (70% of

¹Note that the data set has been de-identified – d_i is useful only for splitting \mathcal{D} into training, testing and validation sets in a patient-wise manner.

the set) normal (healthy) images without polyps. Note that the patients in testing set do not appear in the training/validation sets and vice versa. This abnormality proportion (on the testing set) is commonly defined in other anomaly detection literature [174] [199]. These frames were obtained with the Olympus $\text{\textcircled{R}}$ 190 dual focus colonoscopy.

3.3.2 Methods

Our proposed **ADGAN** is shown in Figure 3.2, and comprises a visual generator, a visual discriminator, a latent generator and a latent discriminator. Defining $\mathbf{z} \sim \mathbb{U}(-1, 1)$, with $\mathbf{z} \in \mathbb{R}^Z$ as the latent variable, the visual generator is defined by

$$\hat{\mathbf{x}} = G_v(\mathbf{z}; \theta_v), \quad (3.1)$$

where θ_v denotes the parameter vector of the generator, and $\hat{\mathbf{x}} : \Omega \rightarrow \mathbb{R}^3$ denotes the generated image. Similarly, the latent generator is defined by

$$\hat{\mathbf{z}} = G_l(\hat{\mathbf{x}}; \theta_l), \quad (3.2)$$

where θ_l is again the parameter vector. During training, $G_v(\cdot)$ from (3.1) generates fake images $\hat{\mathbf{x}}$ given \mathbf{z} to fool the visual discriminator, defined as:

$$r = D_v(\hat{\mathbf{x}}; \gamma_v), \quad (3.3)$$

where γ_v represents the discriminator parameters. The generated image $\hat{\mathbf{x}}$ is then fed to the latent generator in (3.2) to produce a fake latent vector $\hat{\mathbf{z}}$ to fool the latent discriminator, defined as

$$r = D_l(\hat{\mathbf{z}}; \gamma_l), \quad (3.4)$$

where γ_l is the discriminator parameters. The training process follows [83], where we minimise the visual generation loss,

$$\begin{aligned} l_{D_v} &= D_v(\mathbf{x}) - D_v(\hat{\mathbf{x}}) + \lambda(\|\nabla D_v(\hat{\mathbf{x}})\|_2 - 1)^2 \\ l_{G_v} &= -D_v(G_v(\mathbf{z})); \end{aligned} \quad (3.5)$$

and the latent generation loss:

$$\begin{aligned} l_{D_l} &= \log(D_l(\mathbf{z})) + \log(1 - D_l(\hat{\mathbf{z}})) \\ l_{G_l} &= \alpha \log(1 - D_l(G_l(\hat{\mathbf{x}}))). \end{aligned} \quad (3.6)$$

To generate realistic images, we also minimise the mean squared error (MSE) loss between the input and generated latent vectors, as in $l_{MSE} = \beta \|\mathbf{z} - G_l(G_v(\mathbf{z}))\|_2^2$. For training, the hyper-parameters α, β are estimated from the validation set within

the range $[0.1, 10]$. The training process consists of N iterations, where the visual generator is trained for $T < N$ iterations, and then the whole model is trained for $(N - T)$ iterations.

During testing, given a sample \mathbf{x} , a latent vector $\hat{\mathbf{z}}$ is produced with (3.2), which is then fed to the visual generator in (3.1) and the anomaly score is computed with:

$$A(\mathbf{x}) = \|\mathbf{x} - G_v(G_l(\mathbf{x}))\|_2^2. \quad (3.7)$$

Small anomaly score indicates normal samples and high anomaly score generally indicates abnormal samples, suggesting the presence of a polyp (see Fig. 11.1 for a few reconstructions examples produced by ADGAN).

3.4 Experiment

In this section, we validate our proposed ADGAN model using the data set described in Sec. 3.3.1. We compare our performance with other baseline approaches and state-of-the-art methods. We show that our model achieves state-of-the-art area under the ROC curve (AUC) results.

3.4.1 Experimental Setup

We pre-process the original colonoscopy image from $1072 \times 1072 \times 3$ resolution to $64 \times 64 \times 3$ to reduce the computational cost of the training and inference processes. The model selection is done with the validation set mentioned in Sec. 3.3.1. This method is implemented using Pytorch [171] and the code will be publicly available upon acceptance of the paper. We use Adam [57] optimiser during training with a learning rate of 0.0001. Our model has a similar backbone architecture as the other competing methods in Tab. 3.1. In particular, the visual generator and discriminator are based on the improved GAN [63] and use four residual convolution and four residual de-convolution layers, respectively [63]. The latent generator and discriminator are based on DCGAN [55] with three convolution/de-convolution layers. The number of filters per layer for our visual discriminator are (64, 128, 256, 512) (reverse order for visual generator). The number of filters per layer for our latent generator are (64, 128, 256, 512), and we use (256, 128, 64) as the number of filters per layer for our latent discriminator. To train the model, we first train the visual generator and discriminator for 80000 iterations while fixing the parameters of latent generator and discriminator. We then jointly train the whole framework for 20000 iterations, with a batch size of 64.

Methods	AUC
DAE [60]	0.6294
VAE [51]	0.6478
OC-GAN [174]	0.5916
f-AnoGAN(ziz) [199]	0.6376
f-AnoGAN(izi) [199]	0.6638
f-AnoGAN(izif) [199]	0.6913
ADGAN	0.7296

Table 3.1: Comparison between our proposed ADGAN and other state of the art methods.

3.4.2 Anomaly Detection Results

We compare the proposed ADGAN with state-of-the-art approaches, including OC-GAN [174], f-anogan and its variants [199] that involve image-to-image MSE loss (izi), Z-to-Z MSE loss (ziz) and its hybrid version (izif). We also compare our method with some baseline approaches, including deep auto-encoder [60] and variational auto-encoder [51]. The anomaly score $A(\mathbf{x})$ in (3.7) is used to indicate the presence of polyps. For the encoder-decoder architecture comparison, the models adopt similar structure as our latent generator and latent discriminator. For GAN-based methods comparison, we use the same structure as our visual generator and latent discriminator with similar model capacity. We use area under the ROC curve (AUC) as the measurement for performance validation [174, 199]. As shown in Table 3.1, our ADGAN model outperforms other methods.

3.4.3 Image Reconstruction from ADGAN

Figure 11.1 demonstrates the input images (first row) from testing set and their reconstructed images (second row) with our proposed ADGAN model. We manually mark the abnormal polyp lesions using red circles. Our model reconstructs the abnormal input images to their healthy versions, leading to substantial reconstruction error (anomaly score) due to visual differences. The normal images from testing set are generally reconstructed well producing small reconstruction errors (anomaly score).

3.5 Conclusions

In conclusion, we proposed a GAN-based framework (ADGAN) for anomaly detection using one-class learning on a colonoscopy data set. The model was trained end-to-end and experiments show that our model achieved the state-of-the-art anomaly detection

result. We solve the issues of mapping between input image and GAN's latent space using a second GAN model, and proposed a new loss function that combines MSE loss, Wasserstein loss and standard BCE loss. In the future, we plan to extend our model to work with colonoscopy images showing feces and water, as explained in Sec. [3.3.1](#).

Statement of Authorship

Title of Paper	Constrained Contrastive Distribution Learning for Unsupervised Anomaly Detection and Localisation in Medical Images
Publication Status	<input checked="" type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	Published In International Conference on Medical Imaging Computing and Computer-Assisted Intervention (MICCAI) 2021.

Principal Author

Name of Principal Author (Candidate)	Yu Tian		
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper.		
Overall percentage (%)	90		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature	_____	Date	09/03/2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Guansong Pang		
Contribution to the Paper	Discussion and writing the revision.		
Signature	_____	Date	09/03/2022

Name of Co-Author	Fengbei Liu		
Contribution to the Paper	Discussion and writing the revision.		
Signature	_____	Date	09/03/2022

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Yuanhong Chen		
Contribution to the Paper	Discussion and writing the revision.		
Signature	_____	Date	09/03/2022

Name of Co-Author	Seon Ho Shin		
Contribution to the Paper	Discussion and writing the revision.		
Signature	_____	Date	09/03/2022

Name of Co-Author	Rajvinder Singh		
Contribution to the Paper	Discussion and writing the revision.		
Signature	_____	Date	09/03/2022

Name of Co-Author	Johan W Verjans		
Contribution to the Paper	Discussion and writing the revision.		
Signature	_____	Date	09/03/2022

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	Discussion and writing the revision.		
Signature	_____	Date	09/03/2022

Please cut and paste additional co-author panels here as required.

Chapter 4

Constrained Contrastive Distribution Learning for Unsupervised Anomaly Detection and Localisation in Medical Images

Abstract

Unsupervised anomaly detection (UAD) learns one-class classifiers exclusively with normal (i.e., healthy) images to detect any abnormal (i.e., unhealthy) samples that do not conform to the expected normal patterns. UAD has two main advantages over its fully supervised counterpart. Firstly, it is able to directly leverage large datasets available from health screening programs that contain mostly normal image samples, avoiding the costly manual labelling of abnormal samples and the subsequent issues involved in training with extremely class-imbalanced data. Further, UAD approaches can potentially detect and localise any type of lesions that deviate from the normal patterns. One significant challenge faced by UAD methods is how to learn effective low-dimensional image representations to detect and localise subtle abnormalities, generally consisting of small lesions. To address this challenge, we propose a novel self-supervised representation learning method, called Constrained Contrastive Distribution learning for anomaly detection (CCD), which learns fine-grained feature representations by simultaneously predicting the distribution of augmented data and image contexts using contrastive learning with pretext constraints. The learned representations can be leveraged to train more anomaly-sensitive detection models. Extensive experiment results show that our method outperforms current state-of-the-art UAD approaches on three different colonoscopy and fundus screening datasets.

4.1 Introduction

Classifying and localising malignant tissues have been vastly investigated in medical imaging [9, 67, 70, 131, 134, 135, 140, 146, 216]. Such systems are useful in health screening programs that require radiologists to analyse large quantities of images [180, 220], where the majority contain normal (or healthy) cases, and a small minority have abnormal (or unhealthy) cases that can be regarded as anomalies. Hence, to avoid the difficulty of learning from such class-imbalanced training sets and the prohibitive cost of collecting large sets of manually labelled abnormal cases, several papers investigate anomaly detection (AD) with a few or no labels as an alternative to traditional fully supervised imbalanced learning [9, 140, 144, 163, 165, 198, 204, 216, 217, 223]. UAD methods typically train a one-class classifier using data from the normal class only, and anomalies (or abnormal cases) are detected based on the extent the images deviate from the normal class.

Current anomaly detection approaches [32, 34, 56, 139, 198, 216, 229] train deep generative models (e.g., auto-encoder [108], GAN [58]) to reconstruct normal images, and anomalies are detected from the reconstruction error [165]. These approaches rely on a low-dimensional image representation that must be effective at reconstructing normal images, where the main challenge is to detect anomalies that show subtle deviations from normal images, such as with small lesions [216]. Recently, self-supervised methods that learn auxiliary pretext tasks [10, 30, 77, 90, 94, 133] have been shown to learn effective representations for UAD in general computer vision tasks [10, 77, 94], so it is important to investigate if self-supervision can also improve UAD for medical images.

The main challenge for the design of UAD methods for medical imaging resides in how to devise effective pretext tasks. Self-supervised pretext tasks consist of predicting geometric or brightness transformations [10, 77, 94], or contrastive learning [30, 90]. These pretext tasks have been designed to work for downstream classification problems that are not related to anomaly detection, so they may degrade the detection performance of UAD methods [238]. Sohn et al. [208] tackle this issue by using smaller batch sizes than in [30, 90] and a new data augmentation method. However, the use of self-supervised learning in UAD for medical images has not been investigated, to the best of our knowledge. Further, although transformation prediction and contrastive learning show great success in self-supervised feature learning, there are no studies on how to properly combine these two approaches to learn more effective features for UAD.

In this chapter, we propose Constrained Contrastive Distribution learning (CCD), a new self-supervised representation learning designed specifically to learn normality information from exclusively normal training images. The contributions of CCD are: a) contrastive distribution learning, and b) two pretext learning constraints, both of which are customised for anomaly detection (AD). Unlike modern self-supervised

learning (SSL) [30, 90] that focuses on learning generic semantic representations for enabling diverse downstream tasks, CCD instead contrasts the distributions of strongly augmented images (e.g., random permutations). The strongly augmented images resemble some types of abnormal images, so CCD is enforced to learn discriminative normality representations by its contrastive distribution learning. The two pretext learning constraints on augmentation and location prediction are added to learn fine-grained normality representations for the detection of subtle abnormalities. These two unique components result in significantly improved self-supervised AD-oriented representation learning, substantially outperforming previous general-purpose SOTA SSL approaches [10, 30, 77, 94]. Another important contribution of CCD is that it is agnostic to downstream anomaly classifiers. We empirically show that our CCD improves the performance of three diverse anomaly detectors (f-anogan [198], IGD [34], MS-SSIM) [241]). Inspired by IGD [34], we adapt our proposed CCD pretraining on global images and local patches, respectively. Extensive experimental results on three different health screening medical imaging benchmarks, namely, colonoscopy images from two datasets [21, 139], and fundus images for glaucoma detection [121], show that our proposed self-supervised approach enables the production of SOTA anomaly detection and localisation in medical images.

4.2 Method

In this section, we introduce the proposed approach, depicted in the diagram of Fig. 9.1. Specifically, given a training medical image dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$, with all images assumed to be from the normal class and $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{H \times W \times C}$, our approach aims to learn anomaly detection and localisation using three modules: 1) a self-supervised constrained contrastive feature learner that pre-trains an encoding network $f_\theta : \mathcal{X} \rightarrow \mathcal{Z}$ (with $\mathcal{Z} \subset \mathbb{R}^{d_z}$) tailored for anomaly detection, 2) an anomaly classification model $h_\psi : \mathcal{Z} \rightarrow [0, 1]$ that is built upon the pre-trained network, and 3) an anomaly localiser that leverages the classifier $h_\psi(f_\theta(\mathbf{x}_\omega))$ to localise an abnormal image region $\mathbf{x}_\omega \in \mathbb{R}^{\hat{H} \times \hat{W} \times C}$, centred at $\omega \in \Omega$ (Ω is the image lattice) with height $\hat{H} \ll H$ and width $\hat{W} \ll W$. The approach is evaluated on a testing set $\mathcal{T} = \{(\mathbf{x}, y, \mathbf{m})_i\}_{i=1}^{|\mathcal{T}|}$, where $y \in \mathcal{Y} = \{\text{normal, abnormal}\}$, and $\mathbf{m} \in \mathcal{M} \subset \{0, 1\}^{H \times W \times C}$ denotes the segmentation mask of the lesion in the image \mathbf{x} . For adapting our CCD pretraining on patch representations, we simply crop the training images into patches before applying our method.

4.2.1 Constrained Contrastive Distribution Learning

Contrastive learning has been used by self-supervised learning methods to pre-train encoders with data augmentation [30, 90, 238] and contrastive learning loss [207]. The idea is to sample functions from a data augmentation distribution (e.g., geometric and

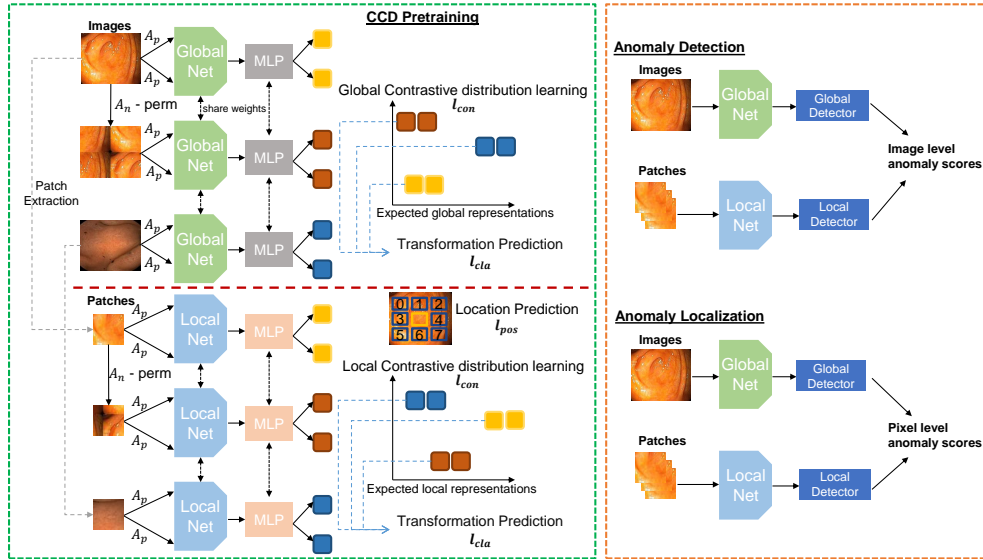


Figure 4.1: Our proposed CCD framework. **Left** shows the proposed pre-training method that unifies a contrastive distribution learning and pretext learning on both global and local perspectives (Sec. 4.2.1), **Right** shows the inference for detection and localisation (Sec. 4.2.2).

brightness transformations), and assume that the same image, under separate augmentations, form one class to be distinguished against all other images in the batch [10, 77]. Another form of pre-training is based on a pretext task, such as solving jigsaw puzzle and predicting geometric and brightness transformations [30, 90]. These self-supervised learning approaches are useful to pre-train classification [30, 90] and segmentation models [162, 252]. Only recently, self-supervised learning using contrastive learning [208] and pretext learning [10, 77] have been shown to be effective in anomaly detection. However, these two approaches are explored separately. In this chapter, we aim at harnessing the power of both approaches to learn more expressive pre-trained features specifically for UAD. To this end, we propose the novel Constrained Contrastive Distribution learning method (CCD).

Contrastive distribution learning is designed to enforce a non-uniform distribution of the representations in the space \mathcal{Z} , which has been associated with more effective anomaly detection performance [208]. Our CCD method constrains the contrastive distribution learning with two pretext learning tasks, with the goal of enforcing further the non-uniform distribution of the representations. The CCD loss is defined as

$$\ell_{CCD}(\mathcal{D}; \theta, \beta, \gamma) = \ell_{con}(\mathcal{D}; \theta) + \ell_{cla}(\mathcal{D}; \beta) + \ell_{pos}(\mathcal{D}; \gamma), \quad (4.1)$$

where $\ell_{con}(\cdot)$ is the contrastive distribution loss, ℓ_{cla} and ℓ_{pos} are two pretext learning

tasks added to constrain the optimisation; and θ , β and γ are trainable parameters. The contrastive distribution learning uses a dataset of **weak data augmentations** $\mathcal{A}_p = \{a_l : \mathcal{X} \rightarrow \mathcal{X}\}_{l=1}^{|\mathcal{A}_p|}$ and **strong data augmentations** $\mathcal{A}_n = \{a_l : \mathcal{X} \rightarrow \mathcal{X}\}_{l=1}^{|\mathcal{A}_n|}$, where $a_l(\mathbf{x})$ denotes a particular data augmentation applied to \mathbf{x} , and the loss is defined as

$$\begin{aligned} \ell_{con}(\mathcal{D}; \theta) = & \\ & - \mathbb{E} \left[\log \frac{\exp \left[\frac{1}{\tau} f_{\theta}(a(\tilde{\mathbf{x}}^j))^{\top} f_{\theta}(a'(\tilde{\mathbf{x}}^j)) \right]}{\exp \left[\frac{1}{\tau} f_{\theta}(a(\tilde{\mathbf{x}}^j))^{\top} f_{\theta}(a'(\tilde{\mathbf{x}}^j)) \right] + \sum_{i=1}^M \exp \left[\frac{1}{\tau} f_{\theta}(a(\tilde{\mathbf{x}}^j))^{\top} f_{\theta}(a'(\tilde{\mathbf{x}}_i^j)) \right]} \right], \end{aligned} \quad (4.2)$$

where the expectation is over $\mathbf{x} \in \mathcal{D}$, $\{\mathbf{x}_i\}_{i=1}^M \subset \mathcal{D} \setminus \{\mathbf{x}\}$, $a(\cdot), a'(\cdot) \in \mathcal{A}_p$, $\tilde{\mathbf{x}}^j = a_j(\mathbf{x})$, $\tilde{\mathbf{x}}_i^j = a_j(\mathbf{x}_i)$, and $a_j(\cdot) \in \mathcal{A}_n$. The images augmented with the functions from the strong set \mathcal{A}_n carry some ‘abnormality’ compared to the original images, which is helpful to learn a non-uniform distribution in the representation space \mathcal{Z} .

We can then constrain further the training to learn more non-uniform representations with a self-supervised classification constraint $\ell_{cla}(\cdot)$ that enforces the model to achieve accurate classification of the strong augmentation function:

$$\ell_{cla}(\mathcal{D}; \beta) = -\mathbb{E}_{\mathbf{x} \in \mathcal{D}, a(\cdot) \in \mathcal{A}_n} \left[\log \mathbf{a}^{\top} f_{\beta}(f_{\theta}(a(\mathbf{x}))) \right], \quad (4.3)$$

where $f_{\beta} : \mathcal{Z} \rightarrow [0, 1]^{|\mathcal{A}_n|}$ is a fully-connected (FC) layer, and $\mathbf{a} \in \{0, 1\}^{|\mathcal{A}_n|}$ is a one-hot vector representing the strong augmentation $a(\cdot) \in \mathcal{A}_n$.

The second constraint is based on the relative patch location from the centre of the training image – this positional information is important for segmentation tasks [110, 162]. This constraint is added to learn fine-grained features and achieve more accurate anomaly localisation. Inspired by [52], the positional constraint predicts the relative position of the paired image patches, with its loss defined as

$$\ell_{pos}(\mathcal{D}; \gamma) = -\mathbb{E}_{\{\mathbf{x}_{\omega_1}, \mathbf{x}_{\omega_2}\} \sim \mathbf{x} \in \mathcal{D}} \left[\log \mathbf{p}^{\top} f_{\gamma}(f_{\theta}(\mathbf{x}_{\omega_1}), f_{\theta}(\mathbf{x}_{\omega_2})) \right], \quad (4.4)$$

where \mathbf{x}_{ω_1} is a randomly selected fixed-size image patch from \mathbf{x} , \mathbf{x}_{ω_2} is another image patch from one of its eight neighbouring patches (as shown in ‘patch location prediction’ in Fig. 9.1), $f_{\gamma} : \mathcal{Z} \times \mathcal{Z} \rightarrow [0, 1]^8$, and $\mathbf{p} = \{0, 1\}^8$ is a one-hot encoding of the synthetic class label.

Overall, the constraints in (5.8) and (5.9) to the contrastive distribution loss in (9.5) are designed to increase the non-uniform representation distribution and to improve the representation discriminability between normal and abnormal samples, compared with [208].

4.2.2 Anomaly Detection and Localisation

Building upon the pre-trained encoder $f_{\theta}(\cdot)$ using the loss in (4.1), we fine-tune two state-of-the-art UAD methods, IGD [34] and F-anoGAN [198], and a baseline method,

multi-scale structural similarity index measure (MS-SSIM)-based auto-encoder [241]. All UAD methods use the same training set \mathcal{D} that contains only normal image samples.

IGD [34] combines three loss functions: 1) two reconstruction losses based on local and global multi-scale structural similarity index measure (MS-SSIM) [241] and mean absolute error (MAE) to train the encoder $f_\theta(\cdot)$ and decoder $g_\phi(\cdot)$, 2) a regularisation loss to train adversarial interpolations from the encoder [16], and 3) an anomaly classification loss to train $h_\psi(\cdot)$. The anomaly detection score of image \mathbf{x} is

$$s_{IGD}(\mathbf{x}) = \xi \ell_{rec}(\mathbf{x}, \tilde{\mathbf{x}}) + (1 - \xi)(1 - h_\psi(f_\theta(\mathbf{x}))), \quad (4.5)$$

where $\tilde{\mathbf{x}} = g_\phi(f_\theta(\mathbf{x}))$, $h_\psi(f_\theta(\mathbf{x})) \in [0, 1]$ returns the likelihood that \mathbf{x} belongs to the normal class, $\xi \in [0, 1]$ is a hyper-parameter, and

$$\ell_{rec}(\mathbf{x}, \tilde{\mathbf{x}}) = \rho \|\mathbf{x} - \tilde{\mathbf{x}}\|_1 + (1 - \rho) (1 - (\nu m_G(\mathbf{x}, \tilde{\mathbf{x}}) + (1 - \nu) m_L(\mathbf{x}, \tilde{\mathbf{x}}))), \quad (4.6)$$

with $\rho, \nu \in [0, 1]$, $m_G(\cdot)$ and $m_L(\cdot)$ denoting the global and local MS-SSIM scores [34]. Anomaly localisation uses (6.21) to compute $s_{IGD}(\mathbf{x}_\omega)$, $\forall \omega \in \Omega$, where $\mathbf{x}_\omega \in \mathbb{R}^{\hat{H} \times \hat{W} \times C}$ is an image region—this forms a heatmap, where large values denote anomalous regions.

F-anoGAN [198] combines generative adversarial networks (GAN) and auto-encoder models to detect anomalies. Training involves the minimisation of reconstruction losses in both the original image and representation spaces to model $f_\theta(\cdot)$ and $g_\phi(\cdot)$. It also uses a GAN loss [58] to model $g_\phi(\cdot)$ and $h_\psi(\cdot)$. Anomaly detection for image \mathbf{x} is

$$s_{FAN}(\mathbf{x}) = \|\mathbf{x} - g_\phi(f_\theta(\mathbf{x}))\| + \kappa \|f_\theta(\mathbf{x}) - f_\theta(g_\phi(f_\theta(\mathbf{x})))\|. \quad (4.7)$$

Anomaly localisation at $\mathbf{x}_\omega \in \mathbb{R}^{\hat{H} \times \hat{W} \times C}$ is achieved by $\|\mathbf{x}_\omega - g_\phi(f_\theta(\mathbf{x}_\omega))\|$, $\forall \omega \in \Omega$.

For the MS-SSIM auto-encoder [241], we train it with the MS-SSIM loss for reconstructing the training images. Anomaly detection for \mathbf{x} is based on $s_{MSI}(\mathbf{x}) = 1 - (\nu m_G(\mathbf{x}, \tilde{\mathbf{x}}) + (1 - \nu) m_L(\mathbf{x}, \tilde{\mathbf{x}}))$, with $\tilde{\mathbf{x}}$ as defined in (6.21). Anomaly localisation is performed with $s_{MSI}(\mathbf{x}_\omega)$ at image regions $\mathbf{x}_\omega \in \mathbb{R}^{\hat{H} \times \hat{W} \times C}$, $\forall \omega \in \Omega$. Inspired by IGD [34], we also pretrain a local model using our CCD pretraining approach based on the local patches for F-anogan [198] and MS-SSIM autoencoder [241], respectively.

4.3 Experiments

4.3.1 Dataset

We test our framework on three health screening datasets. We test both anomaly detection and localisation on the colonoscopy images of Hyper-Kvasir dataset [21]. On the glaucoma datasets using fundus images [121] and colonoscopy dataset [139] that do not have lesion masks, we test anomaly detection only. Detection is assessed with area

under the ROC curve (AUC). Localisation is measured with intersection over union (ioU).

Hyper-Kvasir is a large multi-class public gastrointestinal dataset. The data was collected from the gastroscopy and colonoscopy procedures from Baerum Hospital in Norway. All labels were produced by experienced radiologists. The dataset contains 110,079 images from abnormal (i.e., unhealthy) and normal (i.e., healthy) patients, with 10,662 labelled. We use part of the clean images from the dataset to train our UAD methods. Specifically, 2,100 images from ‘cecum’, ‘ileum’ and ‘bbps-2-3’ are selected as normal, from which we use 1,600 for training and 500 for testing. We also take 1,000 abnormal images and their segmentation masks and stored them in the testing set.

LAG is a large scale fundus image dataset for glaucoma detection [121], containing 4,854 fundus images with 1,711 positive glaucoma scans and 3,143 negative glaucoma scans. We reorganised this dataset for training the UAD methods, with 2,343 normal (negative glaucoma) images for training, and 800 normal images and 1,711 abnormal images with positive glaucoma for testing.

Liu et al.’s colonoscopy dataset is a colonoscopy image dataset for UAD using 18 colonoscopy videos from 15 patients [139]. The training set contains 13,250 normal (healthy) images without any polyps, and the testing set contains 967 images, having 290 abnormal images with polyps and 677 normal (healthy) images without polyps.

4.3.2 Implementation Details

For pre-training, we use Resnet18 [91] as the backbone architecture for the encoder $f_\theta(\mathbf{x})$, and similarly to previous works [30, 208], we add an MLP to this backbone as the projection head for the contrastive learning. All images from the Hyper-Kvasir [21] and LAG [121] datasets are resized to 256×256 pixels. For the Liu et al.’s colonoscopy dataset, images are resized to 64×64 pixels. The batch size is set to 32 and learning rate to 0.01 for the self-supervised pre-training. We investigate the impact of different strong augmentations in \mathcal{A}_n such as rotation, permutation, cutout and Gaussian noise. All weak augmentations in \mathcal{A}_p are the same as SimCLR [30] (i.e., colour jittering, random grey scale, crop, resize, and Gaussian blur). The model is trained using SGD optimiser with temperature 0.2. The encoder $f_\theta(\cdot)$ outputs a 128 dimensional feature in \mathcal{Z} . All datasets are pre-trained for 2,000 epochs.

For the training of IGD [34], F-anoGAN [198] and MS-SSIM auto-encoder [34], we use the hyper-parameters suggested by the respective papers. For localisation, we compute the heatmap based on the localised anomaly scores from IGD, where the final map is obtained by summing the global and local maps. In our experiments, the local map is obtained by considering each 32×32 image patch as a instance and apply our proposed self-supervised learning to it. The global map is computed based on the whole image sized as 256×256 . For F-anoGAN and MS-SSIM auto-encoder, we use the same setup as the IGD, where models based the 256×256 whole image and the

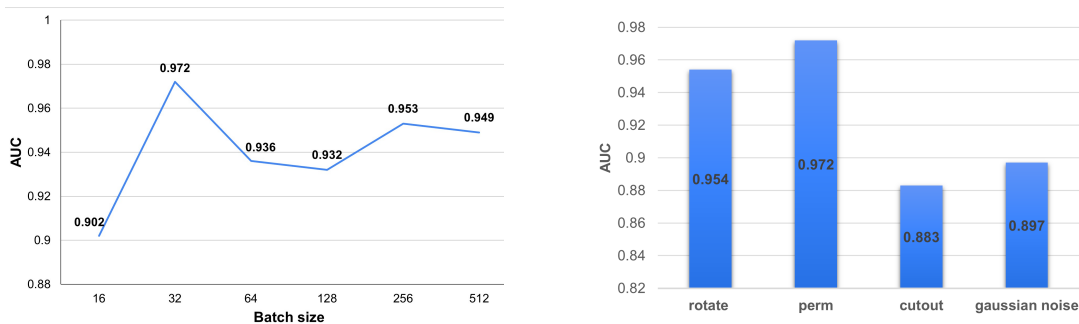


Figure 4.2: **Left:** Anomaly detection performance results based on different batch sizes of self-supervised pre-training. **Right:** Anomaly detection performance in terms of different types of strong augmentations. Both results are on Hyper-Kvasir test set using IGD as anomaly detector.

32×32 patches are trained, respectively. Code will be made publicly available upon paper acceptance.

4.3.3 Ablation Study

In Fig. 4.2 (right), we explore the influence of strong augmentation strategies, represented by rotation, permutation, cutout and Gaussian noise, on the AUC results on Hyper-Kvasir dataset, based on our self-supervised pre-training with IGD as anomaly detector. The experiment indicates that the use of random permutations as strong augmentations yields the best AUC results. We also explore the relation between batch size and AUC results in Fig. 4.2 (left). The results suggest that small batch size (equal to 16) leads to a relatively low AUC, which increases for batch size 32, and then decreases for larger batch sizes. Given these results, we use permutation as the strong augmentation for colonoscopy images and training batch size is set to 32. For the LAG dataset, we omit the results, but we use rotation as the strong augmentation because it produced the largest AUC. We also used batch size of 32 for the LAG dataset.

We also present an ablation study that shows the influence of each loss term in (4.1) in Tab. 11.5, again on Hyper-Kvasir dataset, based on our self-supervised pre-training with IGD. The vanilla contrastive learning in [30, 90] only achieves 91.3% of AUC. After replacing it with our distribution contrastive loss from (9.5), the performance increases by 2.4% AUC. Adding distribution classification and patch position prediction losses boosts the performance by another 2.7% and 0.8% AUC, respectively.

ℓ_{con} [30, 90]	ℓ_{con}	ℓ_{pre}	ℓ_{pat}	AUC - Hyper-Kvasir
✓				0.913
	✓			0.937
	✓	✓		0.964
	✓	✓	✓	0.972

Table 4.1: **Ablation study of the loss terms in (4.1)** on Hyper-Kvasir, using IGD as anomaly detector.

Supervision	Methods	Localisation - IoU
Supervised	U-Net [190]	0.746
	U-Net++ [269]	0.743
	ResUNet [50]	0.793
	SFA [71]	0.611
Unsupervised	RotNet [77]+IGD [34]*	0.276
	CAVGA- R_u [229]	0.349
	Ours - IGD	0.372

Table 4.2: **Anomaly localisation:** Mean IoU results on Hyper-Kvasir on 5 different groups of 100 images with ground truth masks. * indicates that we pretrained the geometric transformation-based anomaly detection [77] using IGD [34] as the UAD method.

4.3.4 Comparison to SOTA Models

In Tab. 5.1, we show the results of anomaly detection on Hyper-Kvasir, Liu et al.’s colonoscopy dataset and LAG datasets. The IGD, F-anoGAN and MS-SSIM methods improve their baselines (without our self-supervision method) from 3.3% to 5.1% of AUC on Hyper-Kvasir, from -0.3% to 12.2% on Liu et al.’s dataset, and from 0.9% to 7.8% on LAG. The IGD with our pre-trained features achieves SOTA anomaly detection AUC on all three datasets. Such results suggest that our self-supervised pre-training can effectively produce good representations for various types of anomaly detectors and datasets. OCGAN [174] constrained the latent space based on two discriminators to force the latent representations of normal data to fall at a bounded area. CAVGA- R_u [229] is a recently proposed approach for anomaly detection and localisation that uses an attention expansion loss to encourage the model to focus on normal object regions in the images. These two methods achieve 81.3% and 92.8% AUC on Hyper-Kvasir, respectively, which are well behind our self-supervised pre-training with IGD of 97.2% AUC.

We also investigate the anomaly localisation performance on Hyper-Kvasir in Tab. 7.3. Compared to the SOTA UAD localisation method, CAVGA- R_u [229], our approach with IGD is more than 3% better in terms of IoU. We also compare our results to **fully supervised methods** [50, 71, 190, 269] to assess how much performance

Methods	Hyper - AUC	Liu et al. - AUC	LAG - AUC
DAE [60]	0.705	0.629 *	-
OCGAN [174]	0.813	0.592 *	-
F-anoGAN [198]	0.907	0.691 *	0.778
ADGAN [140]	0.913	0.730 *	-
CAVGA- R_u [229]	0.928	-	-
MS-SSIM [34]	0.917	0.799	0.823
IGD [34]	0.939	0.787	0.796
RotNet [77]+IGD [34]	0.905	-	-
Ours - MS-SSIM	0.945	0.796	0.839
Ours - F-anoGAN	0.958	0.813	0.787
Ours - IGD	0.972	0.837	0.874

Table 4.3: **Anomaly detection:** AUC results on Hyper-Kvasir, Liu et al.’s colonocopy and LAG, respectively. * indicates that the model does not use Imagenet pre-training.

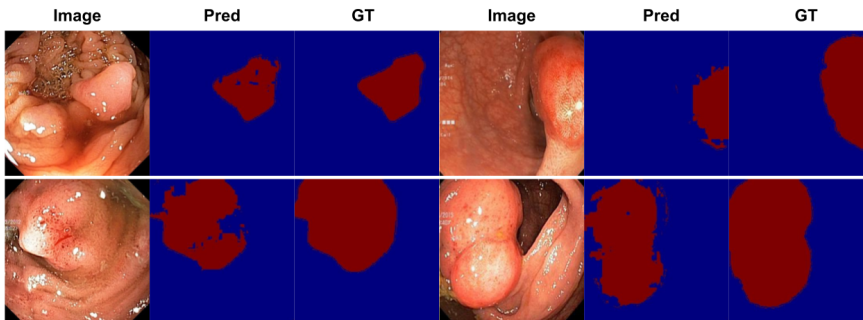


Figure 4.3: Qualitative results of our localisation network based on IGD with self-supervised pre-training on the abnormal images from Hyper Kvasir [21] test set.

is lost by suppressing supervision from abnormal data. The fully supervised baselines [50, 71, 190, 269] use 80% of the annotated 1,000 colonoscopy images containing polyps during training, and 10% for validation and 10% for testing. We validate our approach using the same number of testing samples, but without using abnormal samples for training. The localisation results are post processed by the Connected Component Analysis (CCA) [25]. Notice on Tab. 7.3 that we lose between 0.3 and 0.4 IoU for not using abnormal samples for training.

We present visual anomaly localisation results of our IGD with self-supervised pre-training on the abnormal images from Hyper Kvasir [21] test set in Fig. 4.3. Notice how our model can accurately localise polyps with various size and textures.

4.4 Conclusion

To conclude, we proposed a self-supervised pre-training for UAD named as constrained contrastive distribution learning for anomaly detection. Our approach enforces non-uniform representation distribution by constraining contrastive distribution learning with two pretext tasks. We validate our approach on three medical imaging benchmarks and achieve SOTA anomaly detection and localisation results using three UAD methods. In future work, we will investigate more choices of pretext tasks for UAD.

Statement of Authorship

Title of Paper	Self-supervised Pseudo Multi-class Pre-training for Unsupervised Anomaly Detection and Segmentation in Medical Images
Publication Status	<input type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input checked="" type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	Submitted to Medical Image Analysis

Principal Author

Name of Principal Author (Candidate)	Yu Tian
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper.
Overall percentage (%)	70
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	_____ Date 09/03/2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Fengbei Liu
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper.
Signature	_____ Date 09/03/2022

Name of Co-Author	Guansong Pang
Contribution to the Paper	Discussion and writing the revision.
Signature	_____ Date 09/03/2022

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Yuanhong Chen		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Yuyuan Liu		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Rajvinder Singh		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Johan W Verjans		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Please cut and paste additional co-author panels here as required.

Chapter 5

Self-supervised Pseudo Multi-class Pre-training for Unsupervised Anomaly Detection and Segmentation in Medical Images

Abstract

Unsupervised anomaly detection (UAD) methods are trained with normal (or healthy) images only, but during testing, they are able to classify normal and abnormal (or disease) images. UAD is an important medical image analysis (MIA) method to be applied in disease screening problems because the training sets available for those problems usually contain only normal images. However, the exclusive reliance on normal images may cause the trained UAD model to become over-confident in the normal class classification, and consequently fail in the detection of abnormal cases. Pre-training UAD methods with self-supervised learning, based on computer vision techniques, can mitigate this over-confident normal class classification issue, but they are sub-optimal because they do not explore domain knowledge for designing the pretext tasks, and their contrastive learning losses do not try to cluster the normal training images, which may result in a sparse distribution of normal images that is ineffective for anomaly detection. In this chapter, we propose a new self-supervised pre-training method for MIA UAD applications, named Pseudo Multi-class Strong Augmentation via Contrastive Learning (PMSACL). PMSACL consists of a novel optimisation method that contrasts a normal image class from multiple pseudo classes of synthesised abnormal images, with each class enforced to form a dense cluster in the feature space. In the experiments, we show that our PMSACL pre-training improves the accuracy of SOTA UAD methods on many MIA benchmarks using colonoscopy, fundus screening and Covid-19 Chest X-ray datasets.

5.1 Introduction and Background

Detecting and segmenting abnormal lesions from disease screening datasets is a crucial task in medical images analysis (MIA) [9, 67, 70, 131, 134, 135, 140, 146, 216]. A challenging aspect of this problem is that such screening datasets [180, 220] usually contain a disproportionately large number of normal (or healthy) images, and a tiny amount of abnormal (or disease) images that poorly represent all possible disease sub-classes. Instead of designing a fully supervised training approach to handle such a heavily imbalanced labelled dataset with a poor representation of disease sub-classes, we consider in this chapter an alternative approach based on unsupervised anomaly detection (UAD) [32, 34, 221], which is trained exclusively with normal images. There are two advantages with such UAD strategy: 1) the acquisition of the training set is straightforward given the large proportion of normal images in screening datasets; and 2) it is not necessary to collect a representative training set containing abnormal images from all disease sub-classes. Nevertheless, this UAD strategy is challenging because the model needs to classify abnormal images without being exposed to them during training.

UAD methods are generally based on a one-class classifier (OCC) that learns a normal image distribution from the normal training images, and test image anomalies (or abnormal images) are detected based on the extent that they deviate from the learned distribution [32, 34, 56, 139, 167, 186, 199, 204, 216, 217, 229]. Given their exclusive dependence on normal training images, UAD methods can become over-confident in their classifying, which is an issue that can be mitigated by pre-training the UAD method to solve another classification task. For instance, a common approach is to pre-train the model to classify ImageNet images [221], but there is no guarantee that the learned representations from natural images will be effective for medical images. Another pre-training approach is based on self-supervised learning (SSL) [10, 31, 77, 90, 94, 133, 221], whose effectiveness depends on the relatedness of the pretext tasks and the final MIA classification task, and on the assumptions of the training process. SSL pre-training for UAD methods applied to MIA screening problems have shown promising results [221], but they have been sub-optimally explored given that they were adapted from computer vision methods without using MIA domain knowledge in the design of the pretext tasks or in the training process. For instance, in MIA, normal images should form a single class, while disease images can be divided into sub-classes characterised by variations in the number and appearance of lesions. Instead, previous SSL methods in UAD [208, 214, 221] extend contrastive learning [31] that sub-divides the normal class images into multiple classes formed by geometric or appearance transformations. Such training process is sub-optimal for MIA UAD that needs to discriminate a single tight and dense class of normal images against a relatively small number of abnormal sub-classes that lie outside the normal class distribution.

In this chapter, we propose the Pseudo Multi-class Strong Augmentation via Contrastive

Learning (PMSACL), a new self-supervised pre-training method modelled exclusively with normal training images, and designed to learn effective image representations for different types of downstream UAD methods applied to several MIA problems. The main advantage of PMSACL, compared to previous self-supervised pre-training method for MIA applications [221], is that we rely on MIA domain knowledge to design the training and the pretext tasks. In particular, our training uses contrastive learning to classify training samples into multiple tight and dense clusters in terms of Euclidean distance and cosine similarity, with one cluster representing the normal images and the remaining ones representing pseudo sub-classes of the disease images. These pseudo disease sub-classes are synthesised with our MedMix augmentations that simulate a varying number of lesions of different sizes and appearance in the normal training images (see Fig. 5.2). We summarise our contributions as follows:

- Our PMSACL is the first self-supervised pre-training method specifically designed for MIA UAD applications, where the main advantage lies in the contrastive learning optimisation that learns multiple classes, one for normal images, and the others for pseudo sub-classes of disease images, which are synthesised by our MedMix augmentations by simulating a varying number of lesions of different sizes and appearance.
- We extend our previously published CCD method [221] by proposing two new loss functions to form tighter and denser clusters per class, namely: 1) a multi-centring loss to constrain the feature representations of different classes into a subspace around their class centres; and 2) a non-trivial extension of the normalisation of the standard contrastive loss that repels samples from the same class with less intensity than the samples from different classes.
- The proposed PMSACL is shown to learn effective image representations that can adapt well to different types of downstream UAD methods applied to several MIA problems.

We empirically show that PMSACL pre-training significantly improves the performance of two SOTA anomaly detectors, PaDiM [43] and IGD [34]. Extensive experimental results on four different disease screening medical imaging benchmarks, namely, colonoscopy images from two datasets [21, 139], fundus images for glaucoma detection [121] and Covid-19 Chest X-ray (CXR) dataset [235] show that PMSACL can be used to pre-train diverse SOTA UAD methods to improve their accuracy in detecting and segmenting lesions in diverse medical images.

Relationship to Preliminary Work: An early version of this work was presented in our previously published paper [221]. In this new submission, we considerably expand that previous study by: 1) resolving the strong augmentations issue of CCD that does not synthesise medical image anomalies that are relevant for downstream UAD

applications; 2) addressing the issues around CCD’s contrastive learning that does not consider that the downstream UAD methods will classify one class of normal images and a few sub-classes of disease images; 3) providing a more comprehensive literature review; 4) adding more experiments using datasets from many medical domains; and 5) including a more in-depth analysis of the proposed PMSACL.

5.2 Related Work

UAD approaches [32, 34, 56, 139, 167, 199, 204, 216, 217, 229] can be divided into two categories: predictive-based (e.g., DSVDD [192], OC-SVM [35], and deviation network [167]), and generative-based (e.g., auto-encoder [32, 34, 56, 229] and GAN [3, 140, 199]). Predictive-based UAD approaches train a one-class classifier to describe the distribution of normal data, and discriminate abnormal data using their distance/deviation to the normal data distribution; whereas generative-based UAD approaches train deep generative models to learn latent representations of normal images, and detect anomalies based on image reconstruction error [165]. A fundamental challenge in both types of UAD methods is the learning of expressive feature representations from images, which is particularly important in MIA because abnormal medical images may have subtly looking lesions that can be hard to differentiate from normal images. Hence, if not well trained, these UAD models can become over-confident in classifying normal training data and learn ineffective image representations that will fail to enable the detection and segmentation of lesions.

Pre-training is an effective approach to address the representation challenge described above. A heavily explored pre-training approach is based on using ImageNet [47] pre-trained models, but transferring representations learned from natural images to medical images is not straightforward [221]. Alternatively, the representation challenge can also be tackled by pre-training methods based on self-supervised learning (SSL) that learns auxiliary pretext tasks [10, 31, 77, 90, 94, 133]. SSL is a strategy that has produced effective representations for UAD in general computer vision tasks [10, 77, 94, 214]. However, their application to MIA problems needs to be further investigated because it is not clear how to design effective training or pretext tasks that can work well in the detection of subtle lesions in medical images. Previous UAD methods relied on self-supervised pretext tasks based on the prediction of geometric transformations [10, 77, 94] or contrastive learning using standard data augmentation techniques (e.g., scaling, cropping, etc.) [31, 90] to form a large number of image classes characterising similar and dissimilar pairs. These pretext tasks and training strategy are not specifically related to the detection of subtle anomalies in medical images that contain a normal image class and a small number of disease sub-classes, so they may even degrade the detection accuracy of downstream UAD methods [238].

For SSL UAD pre-training in MIA, the only previous work that we are aware is

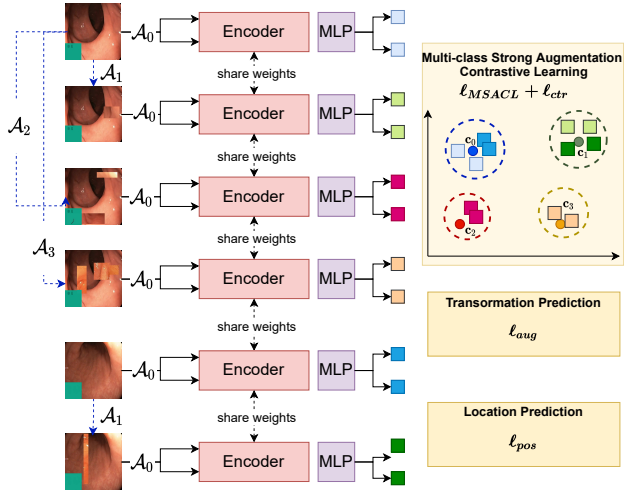


Figure 5.1: **PMSACL**: our proposed self-supervised pre-training for UAD trains four classes of images: the normal images formed by the weak augmentations in distribution \mathcal{A}_0 (blue markers) and three classes of synthesised abnormal images formed by the strong augmentation in distributions $\{\mathcal{A}_n\}_{n=1}^3$ (green, pink and orange markers). The optimisation uses a constrained contrastive learning that trains a four-class classification problem. The different types of strong augmentations are produced by MedMix that introduces a varying number of fake lesions by cutting patches from the normal training images, altering them with random color jittering, Gaussian noise and non-linear intensity transformations, and pasting them to other normal training images.

our previously published CCD method [221] that adapts standard contrastive learning and two general computer vision pretext tasks to image anomaly detection and can be applied to multiple downstream UAD methods. Although achieving good results in many benchmarks, the training explored by CCD does not explore the fact that the downstream UAD methods need to recognise one class of normal images and a small number of sub-classes of disease images, and the CCD’s data augmentation will not synthesise relevant medical image anomalies – both issues can challenge the training of downstream UAD approaches.

5.3 Method

In this section, we introduce the proposed PMSACL pre-training approach depicted in Fig. 9.1. Given a training medical image dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$, with all images assumed to be from the normal class and $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{H \times W \times C}$ (H : height, W : width, C : number of colour channels), our learning strategy involves two stages: 1) the self-

supervised pre-training to learn an encoding network $f_\theta : \mathcal{X} \rightarrow \mathcal{Z}$ (with $\mathcal{Z} \subset \mathbb{R}^Z$), and 2) the fine-tuning of an anomaly detector or segmentation model built from the pre-trained $f_\theta(\cdot)$. The approach is evaluated on a testing set $\mathcal{T} = \{(\mathbf{x}, y, \mathbf{m})_i\}_{i=1}^{|\mathcal{T}|}$, where $y \in \mathcal{Y} = \{\text{normal}, \text{abnormal}\}$, and $\mathbf{m} \in \mathcal{M} \subset \{0, 1\}^{H \times W \times 1}$ denotes the segmentation mask of the lesion in the image \mathbf{x} . Below, we first describe the optimisation proposed for PMSACL in Sec. 5.3.1, then we describe the MedMix data augmentation in Sec. 5.3.2, followed by a brief description of the UAD methods in Sec. 5.3.3.

5.3.1 PMSACL Pre-training

The gist of our proposed PMSACL lies in the idea of discriminating the distribution of weakly augmented samples (simulating normal images) from the distributions of different types of strongly augmented samples (simulating multiple classes of abnormal images). Instead of attracting and repelling samples within and between a large number of image classes [208, 214, 221], we propose a new contrastive loss to separate samples from the normal class and samples from pseudo abnormal sub-classes, and to enforce the clusters representing the normal and abnormal sub-classes to be dense and tight. To this end, our proposed loss is defined as:

$$\ell(\mathcal{D}; \theta, \beta, \gamma) = \ell_{ctr}(\mathcal{D}; \theta) + \ell_{PMSACL}(\mathcal{D}; \theta) + \ell_{aug}(\mathcal{D}; \beta) + \ell_{pos}(\mathcal{D}; \gamma), \quad (5.1)$$

where $\ell_{ctr}(\cdot)$ denotes the new distribution multi-centring loss, $\ell_{PMSACL}(\cdot)$ represents the new PMSACL contrastive loss, $\ell_{aug}(\cdot)$ and $\ell_{pos}(\cdot)$ are the pretext learning losses to regularise the optimisation [221], and θ , β and γ are trainable parameters. The loss terms in (11.2) rely on **weak** data augmentation distribution, denoted by \mathcal{A}_0 , and $|\mathcal{A}|$ **strong** data augmentation distributions, represented by $\{\mathcal{A}_n\}_{n=1}^{|\mathcal{A}|}$, each denoting a different type of augmentation. From each of these distributions, we can sample augmentation functions $a : \mathcal{X} \rightarrow \mathcal{X}$.

The multi-centring loss in (11.2) depends on the estimation of the mean representation for each augmentation distribution, computed as

$$\mathbf{c}_n = \mathbb{E}_{\mathbf{x} \in \mathcal{D}, a \sim \mathcal{A}_n} [f_\theta(a(\mathbf{x}))], \quad (5.2)$$

where $n \in \{0, \dots, |\mathcal{A}|\}$, with \mathbf{c}_n being the mean representation of the training data augmented by the functions sampled from \mathcal{A}_n . Note that these mean representations are computed at the beginning of the training and frozen for the rest of the training. The distribution multi-centring loss is then defined as:

$$\ell_{ctr}(\mathcal{D}; \theta) = \mathbb{E}_{\mathbf{x} \in \mathcal{D}, n \in \{0, \dots, |\mathcal{A}|\}, a \sim \mathcal{A}_n} \|f_\theta(a(\mathbf{x})) - \mathbf{c}_n\|^2, \quad (5.3)$$

which pulls the representations of augmented samples toward their mean representations in (5.2), making the augmentation clusters dense and tight in Euclidean space.

To further enforce the separation between different clusters and the tightness within each cluster, we introduce a novel contrastive learning loss function. In our contrastive learning, we maximise the cosine similarity of samples that belong to the same class (i.e., normal or one of the abnormal sub-classes) and minimise the cosine similarity of samples belonging to different classes. An interesting aspect of this optimisation is that samples are centred by their own cluster mean representation \mathbf{c}_n from (5.2), so our contrastive learning, combined with the multi-centred loss in (5.3) will cluster samples of the same class not only in Euclidean space, but also in inner product space (with cosine measuring similarity between samples). Such re-formulated contrastive learning, combined with the multi-centring loss (5.3), results in a loss that produces multiple clusters, where cluster $n = 0$ contains the normal images and the others, denoted by $n \in \{1, \dots, |\mathcal{A}|\}$, have the synthesised abnormal images. Our PMSACL loss is defined as:

$$\ell_{PMSACL}(\mathcal{D}; \theta) = \mathbb{E}_{\mathbf{x} \in \mathcal{D}, n \in \{0, \dots, |\mathcal{A}|\}, l \in \{0, 1\}} [\ell_{PMSACL}^x(\mathbf{x}^{(n,l)}, \mathcal{D}; \theta)] \quad (5.4)$$

where $\mathbf{x}^{(n,l)} = a(\mathbf{x}^{(n)})$ represents one of two (indexed by $l \in \{0, 1\}$) augmented data obtained from the application of a weak augmentation $a \sim \mathcal{A}_0$ on a strongly augmented data denoted by $\mathbf{x}^{(n)} = a(\mathbf{x})$ with $a \sim \mathcal{A}_n$. In (9.5), we have:

$$\ell_{PMSACL}^x(\mathbf{x}^{(n,l)}, \mathcal{D}; \theta) = -\log \frac{\exp \left[\frac{1}{\tau} (\mathbf{f}^{(n,l)})^\top \mathbf{f}^{(n,(l+1) \bmod 2)} \right]}{\sum_{\substack{\mathbf{x}_j \in \mathcal{D} \\ m \in \{0, \dots, |\mathcal{A}|\} \\ k \in \{0, 1\}}} \mathbb{I}(\mathbf{x}_j^{(m,k)} \neq \mathbf{x}^{(n,l)}) \exp [\kappa(n, m) (\mathbf{f}^{(n,l)})^\top \mathbf{f}^{(m,k)}]}, \quad (5.5)$$

where $\mathbb{I}(\cdot)$ denotes an indicator function, $\mathbf{x}_j^{(m,k)}$ is defined similarly as $\mathbf{x}^{(n,l)}$ in (9.5), $m \in \{0, \dots, |\mathcal{A}|\}$ indexes the set of strong augmentations, and $k \in \{0, 1\}$ indexes one of the two weak augmentations applied to the strongly augmented image. Lastly, to further constrain the normal and strongly augmented data representations in (9.5), our PMSACL loss minimises the distance between samples centred by their representation means computed as:

$$\mathbf{f}^{(n,l)} = \frac{f_\theta(\mathbf{x}^{(n,l)}) - \mathbf{c}_n}{\|f_\theta(\mathbf{x}^{(n,l)}) - \mathbf{c}_n\|_2}, \quad (5.6)$$

where \mathbf{c}_n is defined in (5.2). Also in (9.5) to map the representations from the same distribution into a denser region of the hyper-sphere [37], we propose a temperature scaling strategy defined as:

$$\kappa(n, m) = \begin{cases} 1/(\alpha\tau) & , \text{if } n = m \\ 1/\tau & , \text{otherwise} \end{cases}, \quad (5.7)$$

where α is a scaling factor that controls the shrinkage level of the temperature τ . As a result, Eq. (5.7) alters the temperature for the samples that belong to the same strong augmentation distributions (i.e., when $n = m$) to a smaller value α , which allows smaller amount of repelling strength compared to samples that belongs to strong augmentation distributions (i.e., $n \neq m$). Putting all together, the loss in (9.5) clusters the image representations into hyper-spheres and regions within the hyper-spheres, where each hyper-sphere and region represent a different type of augmentation.

Inspired by [214, 221], we further constrain the training in (11.2) with a self-supervised classification constraint $\ell_{aug}(\cdot)$ that enforces the model to classify the strong augmentation function (Fig. 9.1):

$$\ell_{aug}(\mathcal{D}; \beta) = -\mathbb{E}_{\mathbf{x} \in \mathcal{D}, n \in \{0, \dots, |\mathcal{A}|\}, a \sim \mathcal{A}_n} [\log \mathbf{a}_n^\top f_\beta(f_\theta(a(\mathbf{x})))] , \quad (5.8)$$

where $f_\beta : \mathcal{Z} \rightarrow [0, 1]^{|\mathcal{A}|}$ is a fully-connected (FC) layer, and $\mathbf{a}_n \in \{0, 1\}^{|\mathcal{A}|}$ is a one-hot vector representing the strong augmentation distribution (i.e., $\mathbf{a}_n(j) = 1$ for $j = n$, and $\mathbf{a}_n(j) = 0$ for $j \neq n$).

The final constraint in (11.2) is based on the relative patch location from the centre of the training image and is adapted for local patches. This constraint is added to learn positional and texture characteristics of the image in a self-supervised manner. Inspired by [52], the positional constraint predicts the relative position of the paired image patches, with its loss defined as

$$\ell_{pos}(\mathcal{D}; \gamma) = -\mathbb{E}_{\{\mathbf{x}_{\omega_1}, \mathbf{x}_{\omega_2}\} \sim \mathbf{x} \in \mathcal{D}} [\log \mathbf{p}^\top f_\gamma(f_\theta(\mathbf{x}_{\omega_1}), f_\theta(\mathbf{x}_{\omega_2}))] , \quad (5.9)$$

where \mathbf{x}_{ω_1} is a randomly selected fixed-size image patch from \mathbf{x} , \mathbf{x}_{ω_2} is another image patch from one of its eight neighbouring patches, $\omega_1, \omega_2 \in \Omega$ represents indices to the image lattice, $f_\gamma : \mathcal{Z} \times \mathcal{Z} \rightarrow [0, 1]^8$, and $\mathbf{p} = \{0, 1\}^8$ is a one-hot encoding of the patch location. The constraints in (5.8) and (5.9) are designed to improve training regularisation.

5.3.2 MedMix Augmentation

Our MedMix augmentation is designed to augment medical images to simulate multiple lesions. We target a more effective data augmentation for MIA applications than the computer vision augmentations in [221] (e.g., permutations, rotations) that do not simulate medical image anomalies and may yield poor detection performance by downstream UAD methods. We realise that anomalies in different medical domains (e.g., glaucoma and colon polyps) can be visually different, but a commonality among anomalies is that they are usually represented by an unusual growth of abnormal tissue. Hence, we propose the MedMix augmentation to simulate abnormal tissue with a strong augmentation that ‘‘constructs’’ abnormal lesions by the cutting and pasting (from and to normal images) of small and visually deformed patches. This visual deformation

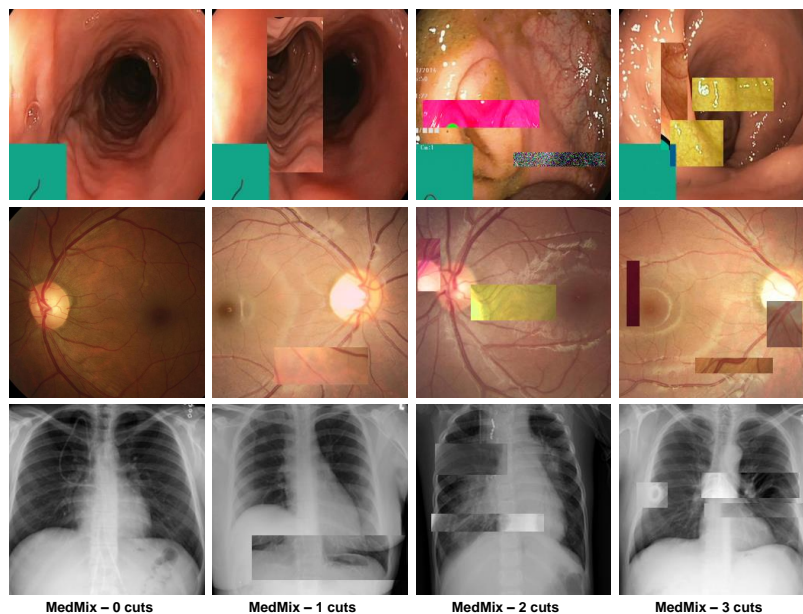


Figure 5.2: Examples of our MedMix data augmentation, showing augmentation \mathcal{A}_0 containing zero synthetic anomalies (leftmost column) and increasingly stronger augmentations $\{\mathcal{A}_n\}_{n=1}^3$ (second to fourth columns) with different number of synthetic anomalies (from one to three).

is achieved by applying other transformations to patches, such as colour jittering, Gaussian noise and non-linear intensity transformations. This approach is inspired by cutmix [256] and CutPaste [120], where our contribution over those approaches is the intensification of the change present in the cropped patches by the appearance transformations above. These transformations are designed to encourage the model to learn abnormalities in terms of localised image appearance, structure, texture and colour.

In practice, we design $|\mathcal{A}| = 4$ strong augmentation distributions, where \mathcal{A}_n includes $n \in \{0, \dots, 3\}$ abnormalities in the image, which means that \mathcal{A}_0 denotes the normal image distribution and $\mathcal{A}_{n \in \{1, 2, 3\}}$ represent the abnormal image distributions, containing $\{1, 2, 3\}$ anomalous regions. Therefore, our loss targets the classification of MedMix augmentations, as shown in Fig. 5.2.

5.3.3 Anomaly Detection and Segmentation

After pre-training $f_\theta(\cdot)$ with PMSACL, we fine-tune it with a SOTA UAD, such as IGD [34] or PaDiM [43]. Those methods use the same training set \mathcal{D} as PMSACL, containing only normal images from healthy patients.

IGD [34] combines three loss functions: 1) two reconstruction losses based on local and global multi-scale structural similarity index measure (MS-SSIM) [241] and mean absolute error (MAE) to train the encoder $f_\theta : \mathcal{X} \rightarrow \mathcal{Z}$ and decoder $g_\phi : \mathcal{Z} \rightarrow \mathcal{X}$, 2) a regularisation loss to train adversarial interpolations from the encoder [16], and 3) an anomaly classification loss to train $h_\psi : \mathcal{Z} \rightarrow [0, 1]$. The anomaly detection score of image \mathbf{x} is defined by

$$s_{IGD}(\mathbf{x}) = \xi \ell_{rec}(\mathbf{x}, \tilde{\mathbf{x}}) + (1 - \xi)(1 - h_\psi(f_\theta(\mathbf{x}))), \quad (5.10)$$

where $\tilde{\mathbf{x}} = g_\phi(f_\theta(\mathbf{x}))$, $h_\psi(\cdot)$ returns the likelihood that \mathbf{x} is a normal image, $\xi \in [0, 1]$ is a hyper-parameter, and

$$\ell_{rec}(\mathbf{x}, \tilde{\mathbf{x}}) = \rho \|\mathbf{x} - \tilde{\mathbf{x}}\|_1 + (1 - \rho) (1 - (\nu m_G(\mathbf{x}, \tilde{\mathbf{x}}) + (1 - \nu) m_L(\mathbf{x}, \tilde{\mathbf{x}}))), \quad (5.11)$$

with $\rho, \nu \in [0, 1]$, $m_G(\cdot)$ and $m_L(\cdot)$ denoting the global and local MS-SSIM scores from the global and local models, respectively [34]. Anomaly segmentation uses (6.21) to compute $s_{IGD}(\mathbf{x}_\omega)$, $\forall \omega \in \Omega$ using global and local models, where $\mathbf{x}_\omega \in \mathbb{R}^{\hat{H} \times \hat{W} \times C}$ is an image patch. This forms a heatmap, where large values of $s_{IGD}(\cdot)$ denote anomalous regions. The final heatmap is formed by summing up the global and local heatmaps.

PaDiM [43] utilises the multi-layer features from the pre-trained network $f_\theta(\cdot)$ to learn a position dependent multi-variate Gaussian distribution of normal image patches. Training uses samples collected from the concatenation of the multi-layer features from each patch position $\omega \in \Omega$ to learn the mean and covariance of the Gaussian model denoted by $\mathcal{N}(\mu_\omega, \Sigma_\omega)$ [43]. Anomaly detection is based on the Mahalanobis distance between the concatenated testing patch feature \mathbf{x}_ω and the learned Gaussian distribution $\mathcal{N}(\mu_\omega, \Sigma_\omega)$ at that patch position $\omega \in \Omega$ to provide a score of each patch position [43]. In particular, anomaly segmentation is inferred using the following anomaly score map:

$$s_{PaDiM}(\mathbf{x}_\omega) = \sqrt{(\mathbf{x}_\omega - \mu_\omega)^\top \Sigma_\omega^{-1} (\mathbf{x}_\omega - \mu_\omega)}, \quad (5.12)$$

and the final score of the whole image \mathbf{x} is defined as: $s_{PaDiM}(\mathbf{x}) = \max_{\omega \in \Omega} s_{PaDiM}(\mathbf{x}_\omega)$.

5.4 Experiments

5.4.1 Datasets

We test our self-supervised pre-training PMSACL on four health screening datasets, where we run experiments for both anomaly detection and localisation. The datasets for anomaly detection and localisation are: the colonoscopy images of Hyper-Kvasir dataset [21], and the glaucoma dataset using fundus images [121]. We also run anomaly detection without localisation experiments on the following datasets: Liu et al.’s colonoscopy

dataset [139], and Covid-19 chest ray dataset [235] – these two datasets do not have lesion segmentation annotations, so we test anomaly detection only.

Hyper-Kvasir is a large multi-class public gastrointestinal imaging dataset [21]. We use a subset of the normal (i.e., healthy) images from the dataset for training. Specifically, 2,100 images from ‘cecum’, ‘ileum’ and ‘bbps-2-3’ are selected as normal, from which we use 1,600 for training and 500 for testing. We also take 1,000 abnormal images and their segmentation masks of polyps to be used exclusively for testing, where all images have size 300×300 pixels.

LAG is a large scale fundus image dataset for glaucoma diagnosis [121]. For the experiments, we use 2,343 normal (negative glaucoma) images for training, and 800 normal images and 1,711 abnormal images with positive glaucoma with annotated attention maps by ophthalmologists for testing, where images are 500×500 pixels. The annotated attention maps are based on eye tracking, in which the maps are used by the ophthalmologists to explore the region of interest for glaucoma diagnosis [121].

Liu et al.’s colonoscopy dataset is a colonoscopy image dataset with 18 colonoscopy videos from 15 patients [139]. The training set contains 13,250 normal (healthy) images without polyps, and the testing set contains 967 images, with 290 abnormal images with polyps and 677 normal (healthy) images without polyps, where all images have size 64×64 pixels.

Covid-X [235] has a training set with 1,670 Covid-19 positive and 13,794 Covid-19 negative CXR images. The test set contains 400 CXR images, consisting of 200 positive and 200 negative images. We train the methods with the 13,794 Covid-19 negative CXR training images and test on the 400 CXR images, where images are 299×299 pixels.

5.4.2 Implementation Details

For the proposed PMSACL pre-training, we use Resnet18 [91] as the backbone architecture for the encoder $f_\theta(\mathbf{x})$, and similarly to previous works [31, 208], we add an MLP to this backbone as the projection head for the contrastive learning, which outputs features in \mathcal{Z} of size 128. All images from the Hyper-Kvasir [21], LAG [121] and Covid-X [235] datasets are resized to 256×256 pixels. For the Liu et al.’s colonoscopy dataset [139], we use the original image size of 64×64 pixels. The batch size is set to 32 and learning rate to 0.01 for the self-supervised pre-training on all datasets. The model is trained using stochastic gradient descent (SGD) optimiser with momentum.

We investigate the impact of different strong augmentations in \mathcal{A}_n , including rotation, permutation, cutout, Gaussian noise and our proposed MedMix. For MedMix patches, we randomly apply colour jittering, Gaussian noise and non-linear intensity transformations (i.e., fisheye and horizontal wave transformations). The weak augmentations in \mathcal{A}_0 are the same as in SimCLR [31], namely: colour jittering, random grey scale, crop, resize, and Gaussian blur.

The model pre-trained with PMSACL is fine-tuned with IGD [34] or PaDiM [43]. For IGD [34], we pre-train the global and local models (see Figure 9.1), where the patch position prediction loss in Eq. 5.9 is only fine-tuned for the local model. For PaDiM [43], we pre-train the global model and use it to fine-tune the anomaly detection and segmentation models. For the training of IGD [34] and PaDiM [43], we use the hyper-parameters suggested by the respective papers. In our experiments, the local map for IGD is obtained by considering each 32×32 -pixel patch as an instance and apply our proposed self-supervised learning to it. The global map for IGD is computed based on the whole image sized as 256×256 pixels for Hyper-Kvasir, LAG and Covid-X datasets. For Liu et al.’s colonoscopy dataset, we only train the model globally with the image size 64×64 . For the auto-encoder in IGD, we use the setup suggested in [34], where the global model is trained with images of size 256×256 pixels or 64×64 for Liu et al.’s colonoscopy dataset, and the local model is trained with image patches of size 32×32 . For PaDiM [43], we only use the default setup in their work and compute the segmentation mask based on the images of size 256×256 pixels for Hyper-Kvasir, LAG and Covid-X datasets, and 64×64 for Liu et al.’s colonoscopy dataset.

5.4.3 Evaluation Measures

The anomaly detection performance is quantitatively assessed by the area under the receiver operating characteristic curve (AUROC), specificity, sensitivity and accuracy. AUROC assesses anomaly detection by varying the classification threshold and computing the area under the ROC curve. Sensitivity and specificity reflect the percentage of positives and negatives that are correctly detected. Accuracy shows the overall performance of correctly detected samples for both positive and negative images, where the classification threshold is estimated with a small validation set that contains 50 normal and 50 abnormal images that are randomly sampled from the testing set. Note that the validation set is only used for threshold estimation. For anomaly segmentation, the performance is measured by Intersection over Union (IoU), Dice score, and Pro-score [11]. IoU is computed by dividing the intersection by the union between the predicted segmentation and the ground truth mask. Dice also takes the predicted segmentation and the ground truth mask and divides two times their intersection by their sum. Pro-score weights the ground-truth masks of different sizes equally [11] to verify if both large and small abnormal lesions are accurately segmented.

5.4.4 Anomaly Detection Results

In this section, we show the anomaly detection results on all datasets.

Methods	AUC	Specificity	Sensitivity	Accuracy
DAE [60]	0.705	0.522	0.756	0.693
OCGAN [174]	0.813	0.691	0.811	0.795
F-Anogan [199]	0.907	0.846	0.915	0.883
ADGAN [140]	0.913	0.879	0.946	0.893
MS-SSIM [34]	0.917	0.857	0.925	0.912
PANDA [185]	0.937	0.805	0.919	0.917
CutPaste [120]	0.949	0.847	0.957	0.932
PaDiM [43]	0.943	0.846	0.929	0.898
CCD - PaDiM	0.978	0.923	0.961	0.967
PMSACL - PaDiM	0.996	0.966	0.981	0.983
IGD [34]	0.939	0.858	0.913	0.906
CCD - IGD	0.972	0.934	0.947	0.956
PMSACL - IGD	0.995	0.947	0.965	0.972

Table 5.1: **Anomaly detection** testing results on **Hyper-Kvasir** in terms of AUC, Specificity, Sensitivity and Accuracy. Best results are highlighted.

Hyper-Kvasir

In Table 5.1, we show the results of anomaly detection on Hyper-Kvasir dataset, where we present results from baseline UAD methods, including OCGAN [174], F-Anogan [199], ADGAN [139], and deep autoencoder (DAE) [60] and its variant with MS-SSIM loss [34]. As discussed in Section 5.3.3, we choose IGD [34] and PaDiM [43] as the anomaly detector for evaluating our proposed PMSACL pre-training approach and compare it with our previously proposed CCD pre-training approach [221] to fine-tune IGD and PaDiM.

Comparing with the baseline UAD methods, the performance of PaDiM and IGD are improved using our PMSACL pre-trained encoder by around 5% and 6% AUC, which achieves SOTA anomaly detection AUC results of 99.6% and 99.5%, respectively, on Hyper-Kvasir. Comparing with our previously proposed CCD pre-training [221], our proposed PMSACL pre-training improves the performance by 2.3% and 1.8% for PaDiM and IGD. This shows that our proposed MedMix and PMSACL loss improve the generalisation ability of the fine-tuning stage for anomaly detection and produce better constrained feature space of normal samples. Moreover, achieving SOTA results on two different types of anomaly detectors suggests that our self-supervised pre-training can produce good representations for both generative and predictive anomaly detectors.

OCGAN [174] constrains the latent space based on two discriminators to force the latent representations of normal data to fall at a bounded area. F-Anogan [199] uses an encoder to extract the feature representations of a input image and use a GAN to reconstruct it. ADGAN [140] uses two generators and two discriminators to produce realistic reconstruction of normal samples. These three methods achieve 81.3%, 90.7% and 91.3% AUC on Hyper-Kvasir, respectively, which are well below our self-supervised

Methods	AUC	Specificity	Sensitivity	Accuracy
MS-SSIM [34]	0.823	0.257	0.937	0.774
F-Anogan [199]	0.778	0.565	0.899	0.763
PANDA [185]	0.789	0.624	0.869	0.767
CutPaste [120]	0.745	0.372	0.788	0.685
PaDiM [43]	0.688	0.314	0.809	0.673
CCD - PaDiM	0.728	0.429	0.779	0.694
PMSACL - PaDiM	0.761	0.466	0.877	0.753
IGD [34]	0.796	0.396	0.958	0.805
CCD - IGD	0.874	0.572	0.944	0.875
PMSACL - IGD	0.908	0.531	0.979	0.884

Table 5.2: **Anomaly detection** testing results on **LAG** in terms of AUC, Specificity, Sensitivity, Precision and Recall. Best results are highlighted.

PMSACL pre-training with IGD and PaDiM. Also, the recently proposed state-of-the-art (SOTA) methods PANDA [185] and CutPaste [120] achieve significantly inferior performance than our PMSACL pre-trained anomaly detectors. Note that CutPaste uses a similar augmentation strategy as MedMix, but with inferior results, indicating the effectiveness of our proposed PMSACL self-supervised loss function. Furthermore, PaDiM with our PMSACL pre-training can achieve the SOTA results of 96.6% specificity, 98.1% sensitivity and 98.3% accuracy. It improves the previous PaDiM using CCD pre-training by 4.3%, 2% and 1.6% for these three evaluation measures. Finally, PaDiM pre-trained with PMSACL significantly outperforms the PaDiM pre-trained with ImageNet [43] by 12%, 5.2% and 8.5% in terms of these three evaluation measures.

LAG

We evaluate the performance of our PMSACL pre-training on the LAG dataset and show results on Table 5.2. Our PMSACL pre-training improves PaDiM and IGD AUCs by 7.3% and 11.2%, compared with their ImageNet pre-trained model, where the PMSACL pre-trained IGD achieves the SOTA results of 90.8% AUC, 97.9% sensitivity and 88.4% accuracy. Comparing with our previous CCD pre-trained PaDiM and IGD [221], our proposed PMSACL pre-trained PaDiM and IGD surpass them by 3.3% and 3.4% in terms of AUC. The MS-SSIM autoencoder [34], F-Anogan [199], PANDA [185], and CutPaste [120] baselines achieve 82.3%, 77.8%, 78.9%, and 74.5% AUC, respectively, which are significantly inferior compared with our PMSACL pre-trained IGD. For LAG, IGD with both reconstruction and anomaly classification constraints can generally outperform PaDiM variants, indicating the superiority of IGD when handling the subtle image features to detect glaucoma.

Methods	AUC	Specificity	Sensitivity	Accuracy
DAE [60]	0.629*	0.733*	0.554*	0.597*
OCGAN [174]	0.592*	0.716*	0.534*	0.624*
ADGAN [140]	0.730*	0.852*	0.496*	0.713*
F-Anogan [199]	0.735	0.865	0.579	0.694
PANDA [185]	0.719	0.846	0.551	0.671
CutPaste [120]	0.779	0.895	0.772	0.738
PaDiM [43]	0.741	0.851	0.738	0.751
CCD - PaDiM	0.789	0.946	0.792	0.767
PMSACL - PaDiM	0.814	0.973	0.725	0.803
IGD [34]	0.787	0.914	0.596	0.743
CCD - IGD	0.837	0.985	0.774	0.815
PMSACL - IGD	0.851	0.986	0.792	0.829

Table 5.3: **Anomaly detection** testing results on **Liu et al.’s colonoscopy** in terms of AUC, Specificity, Sensitivity and Accuracy. * indicates that the model does not use ImageNet pre-training. Best results are highlighted.

Liu et al.’s Colonoscopy Dataset

We further test our approach on Liu et al.’s colonoscopy dataset [140], as shown in Table 5.3. Our PMSACL pre-trained PaDiM improves the ImageNet pre-trained PaDiM by 7.3% AUC, and CCD pre-trained PaDiM by 2.5% of AUC. The IGD with the PMSACL pre-trained encoder achieves the SOTA result of 85.1% AUC, surpassing the previous CCD and ImageNet pre-trained IGD by 1.4% and 6.4% AUC, respectively.

Compared with other UAD approaches, such as F-Anogan, ADGAN, OCGAN, PANDA, and CutPaste that achieve 73.5%, 73%, 59.2%, 70.2% and 77.9% AUC, our PMSACL pre-trained IGD and PaDiM produce substantially better results. The gap between PaDiM and IGD may be due to the low resolution of the images in this dataset, which hinders the PaDiM performance that requires dense intermediate feature maps. The additional results of the PMSACL pre-trained IGD are specificity of 98.6%, sensitivity of 79.2%, and accuracy of 82.9%, which demonstrate the robustness of our proposed model.

Covid-X

Table 5.4 shows that Covid-X results, where our PMSACL pre-trained PaDiM and IGD methods achieve 65.8% and 87.2% AUC on the Covid-X dataset, significantly surpassing their ImageNet pre-trained approaches by 4.4% and 17.2% AUC, and CCD pre-trained by 2.6% and 12.6% AUC. Moreover, our approaches achieve significantly better performance when compared against current SOTA MS-SSIM, F-Anogan, PANDA, and CutPaste baselines. The small abnormal lesions in chest X-ray images are hard to detect, so the generative-based anomaly detector IGD can learn more effectively

Methods	AUC	Specificity	Sensitivity	Accuracy
MS-SSIM [34]	0.634	0.572	0.406	0.577
F-Anogan [199]	0.669	0.718	0.365	0.532
PANDA [185]	0.629	0.762	0.447	0.591
CutPaste [120]	0.658	0.701	0.494	0.648
PaDiM [43]	0.614	0.753	0.318	0.559
CCD - PaDiM	0.632	0.673	0.569	0.616
PMSACL - PaDiM	0.658	0.749	0.467	0.615
IGD [34]	0.699	0.885	0.490	0.688
CCD - IGD	0.746	0.851	0.595	0.722
PMSACL - IGD	0.872	0.863	0.775	0.813

Table 5.4: **Anomaly detection** testing results on **Covid-X** in terms of AUC, Specificity, Sensitivity and Accuracy, respectively. Best results are highlighted.

Dataset	AUC	Specificity	Sensitivity	Accuracy
Hyper-Kvasir [21]	0.0084	0.0079	0.0127	0.0036
LAG [121]	0.0163	0.0085	0.0105	0.0121
Covid-X [235]	0.0107	0.0149	0.0092	0.0171

Table 5.5: The standard deviation of five-run experimental results on the Hyper-Kvasir, LAG and Covid-X based on the PMSACL pre-trained PaDiM anomaly detector. This results should be studied together with the results shown in Tables 5.1, 5.2, 5.4.

the fine-grained appearances of normal images, leading to better ability to detect unseen anomalous regions during testing with the SOTA results of 87.2% AUC, 77.5% of sensitivity and 81.3% of accuracy. The PMSACL pre-trained IGD achieves 86.3% specificity, which is competitive with the result from IGD pre-trained with ImageNet. It can also be observed that our PMSACL pre-trained PaDiM and IGD improve sensitivity by 14.9% and 28.5%, when compared to the ImageNet pre-trained PaDiM and IGD.

Variability in the Results

We show on Table 5.5 the standard deviation computed from the AUC, specificity, sensitivity and accuracy results of five different trainings based on different model initialisation of the PMSACL pre-trained PaDiM detector. These results in Table 5.5 should be studied together with the Tables 5.1, 5.2, 5.4. In general, we conclude that the differences between the methods described in the sections above can be considered significant in most cases given that the standard deviation only varies from 0.5% to 1.5%.

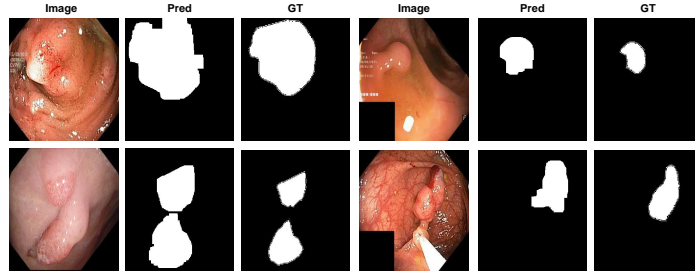


Figure 5.3: Localisation of four abnormal images from Hyper Kvasir [121], with their predictions (Pred) and ground truth annotations (GT), using PaDiM with PMSACL pre-training.

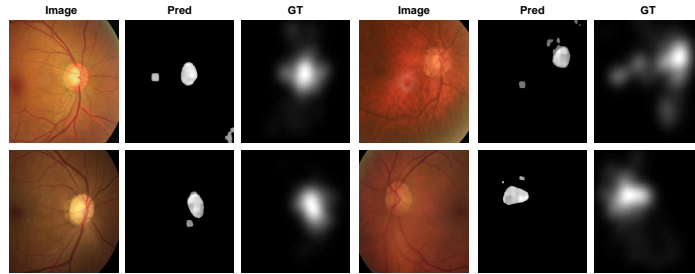


Figure 5.4: Localisation of four abnormal images from LAG [121], with their predictions (Pred) and ground truth attention maps (GT), using IGD with PMSACL pre-training.

5.4.5 Anomaly Localisation Results

In this section, we show the anomaly localisation results on Hyper-Kvasir and LAG.

Hyper-Kvasir

We demonstrate the anomaly localisation performance on Hyper-Kvasir on Table 7.3. Following [221], we randomly sample 100 abnormal images from the test set and compute the mean segmentation performance over five different such groups of 100 images. The proposed PMSACL pre-training improves the IGD and PaDiM by 1.2% and 2.8% IoU compared with the CCD pre-training, and 8.1% and 6.4% IoU with respect to the ImageNet pre-training, respectively. In addition, our PMSACL pre-trained PaDiM shows the SOTA result of 40.6% IoU and 55.4% Dice, demonstrating the effectiveness of our PMSACL approach for abnormal lesion segmentation. The CCD version of PaDiM achieves the SOTA result of 88.1% Pro-score.

Methods	IoU	Dice	Pro
PaDiM [43]	0.341	0.475	0.803
CCD - PaDiM	0.378	0.497	0.881
PMSACL - PaDiM	0.406	0.554	0.854
IGD [34]	0.303	0.417	0.794
CCD - IGD	0.372	0.502	0.865
PMSACL - IGD	0.384	0.521	0.876

Table 5.6: **Anomaly localisation:** Mean IoU, Dice and PRO-AUC testing results on **Hyper-Kvasir** on 5 different groups of 100 images with ground truth masks. Best results for each case are highlighted.

Methods	IoU	Dice	Pro
PaDiM [43]	0.427	0.579	0.596
CCD - PaDiM	0.462	0.612	0.634
PMSACL - PaDiM	0.475	0.643	0.628
IGD [34]	0.409	0.539	0.603
CCD - IGD	0.509	0.645	0.677
PMSACL - IGD	0.516	0.667	0.693

Table 5.7: **Anomaly localisation:** Mean IoU, Dice and Pro-AUC testing results on abnormal samples from **LAG** test set. Best results are highlighted.

LAG

We further demonstrate the segmentation results on LAG dataset on Table 5.7. The PMSACL pre-trained IGD achieves the SOTA result of 51.6% IoU, 66.7% Dice and 69.3% Pro-score, showing that our model can effectively segment different types of lesions, such as colon polyps or optic disk and cup with Glaucoma. Moreover, PaDiM pre-trained with PMSACL improves PaDiM pre-trained with CCD and ImageNet by 1.3% and 4.8% IoU, respectively. Also, PaDiM with PMSACL pre-training achieves 64.3% Dice and 62.8% Pro-score, which are comparable to the SOTA results by the PMSACL pre-trained IGD.

5.4.6 Qualitative Results

In this section, we show examples of anomaly localisation and detection results, and t-SNE results displaying the distribution of image representations of the normal and pseudo abnormal classes in the feature space.

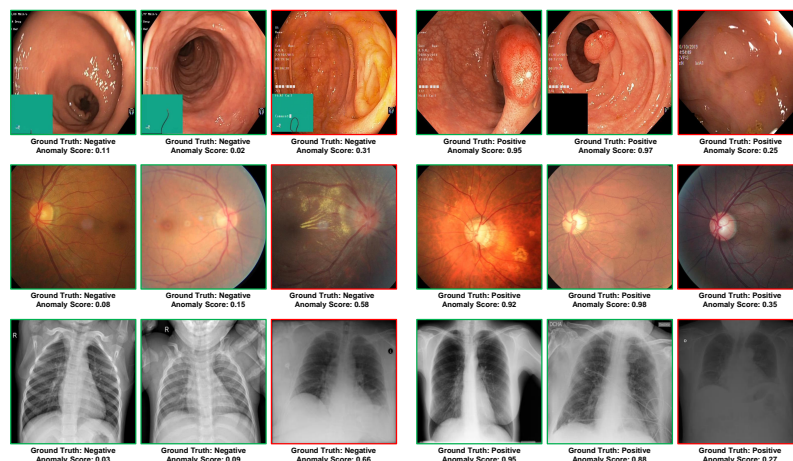


Figure 5.5: Visual detection results and anomaly scores produced by the PMSACL pre-trained IGD on three different datasets: Hyper-Kvasir (top), LAG (middle), Covid-X (bottom). Anomaly scores > 0.5 classifies the image as positive, otherwise, the image is classified as negative. Correctly classified images are marked with green boxes, and incorrectly classified cases are marked with red boxes.

Anomaly Localisation and Detection Visual Results

The visualisation of polyp localisation results of PaDiM with PMSACL pre-training on Hyper-Kvasir [21] is shown in Fig. 5.3. Notice that our model can effectively localise colon polyps with various sizes and shapes. We also show the localisation results based on the pixel-level anomaly scores of IGD with PMSACL pre-training on the LAG dataset in Fig. 5.4.

The visual anomaly detection results of IGD pre-trained with PMSACL on the Hyper-Kvasir [21] test set is shown in Figure 5.5.

Visualisation of t-SNE results

To validate our proposed PMSACL pre-training, we show a comparison of the image representations produced by ImageNet, CCD, DROC and PMSACL pre-training, using t-SNE on Hyper-Kvasir testing images, in Fig. 5.6. The proposed PMSACL appears to cluster all the normal data into a denser and tighter region of the representation space, where the abnormal data fall outside of this region in relatively distinct three clusters. In contrast, the models pre-trained with the other approaches produce a poorly clustered normal data that is likely to challenge the training of the downstream UAD method.

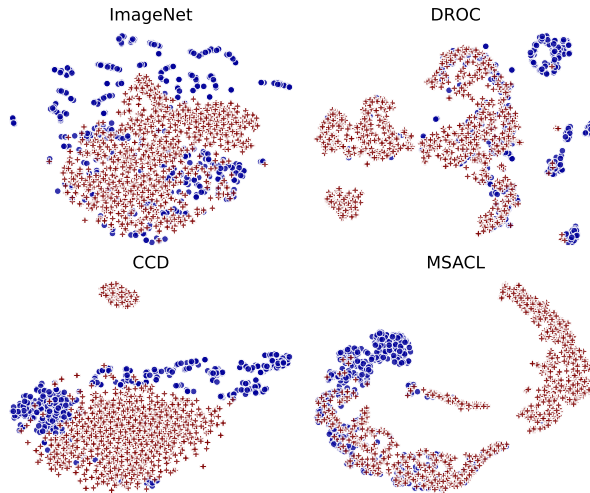


Figure 5.6: t-SNE results of the image representations of the test set of Hyper-Kvasir [21] learned by IGD [34] after being pre-trained on ImageNet [47], or self-supervised with DROC [208], CCD [221], and our PMSACL. Compared to other methods, PMSACL clusters the normal image representations (blue points) in a tighter and denser region, and separates anomalous representations into three clusters (red points), which can be associated with the three classes of synthesised abnormal images formed by simulating a varying number of lesions of different sizes and appearance in the normal images.

5.4.7 Ablation Study

In this section, we study the roles played by PMSACL components. We start by investigating the loss terms in (11.2). Then we study the impact of using different types of strong data augmentation to generate the pseudo abnormal images and the number of pseudo abnormal classes in MedMix (i.e., the size $|\mathcal{A}|$ in (11.2)). We also compare our approach with other recently proposed self-supervised pre-training approaches.

Loss Terms in PMSACL pre-training

We present an ablation study that shows the influence of each term of our proposed PMSACL pre-training in (11.2) following PaDiM fine-tuning in Table 11.5 on Hyper-Kvasir and LAG datasets. Starting from ℓ_{PMSACL} without temperature scaling strategy $\kappa(n, m)$ and multi-centring loss ℓ_{ctr} , we notice that the performance can reach 94.9% and 72.5% AUC on on Hyper-Kvasir and LAG datasets, respectively. Adding the multi-centring loss ℓ_{ctr} provides an improvement of $\approx 2\%$ AUC. Then, adding the temperature scaling $\kappa(n, m)$ from (5.7) provides $\approx 0.8\%$ improvement. This indicate that the joint training between ℓ_{PMSACL} and ℓ_{ctr} with temperature scaling strategy

ℓ_{PMSACL}	ℓ_{ctr}	$\kappa(n, m)$	ℓ_{aug}	ℓ_{pos}	AUC - Hyper	AUC - LAG
✓					0.949	0.717
✓	✓				0.971	0.738
✓	✓	✓			0.979	0.745
✓	✓	✓	✓		0.991	0.756
✓	✓	✓	✓	✓	0.996	0.761

Table 5.8: **Ablation study of the PMSACL loss terms** on the test sets of Hyper-Kvasir and LAG, using PaDiM [43] as anomaly detector, with our MedMix as strong augmentations.

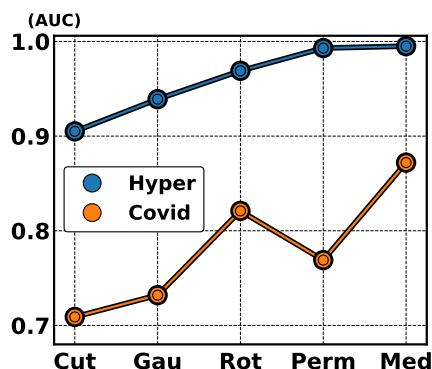


Figure 5.7: Anomaly detection testing results in terms of different types of strong augmentations (i.e., Cutmix, Gaussian noise, Rotation, Permutation, and our MedMix) on Hyper-Kvasir and Covid-X, where our PMSACL is used as self-supervised pre-training, and IGD [34] is used as the anomaly detector.

can learn better fine-grained low-dimensional features for the downstream anomaly detectors (i.e., producing denser and tighter cluster for normal images). Finally, adding ℓ_{aug} and ℓ_{pos} further boost the performance to reach 99.6% and 76.1% AUC on both datasets.

Strong Augmentations

In Fig. 5.7, we explore the influence of strong augmentation strategies, represented by rotation, permutation, cutout, Gaussian noise and our proposed MedMix on the AUC results of Hyper-Kvasir and Covid-X datasets, based on our self-supervised PMSACL pre-training with IGD as anomaly detector. The performance of our MedMix reaches the SOTA results of 99.5% and 87.2% on those datasets. The second best AUC (96.9%) on Hyper-Kvasir uses random permutations, which were used in CCD pre-training [221], producing an AUC 0.2% worse than our MedMix. For Covid-X,

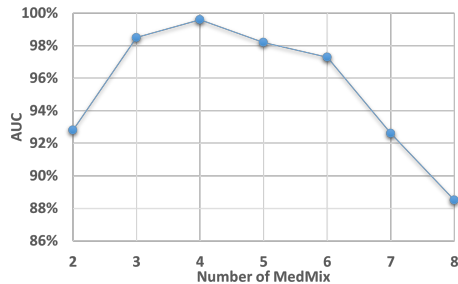


Figure 5.8: Influence of the number of MedMix augmentation distributions $|\mathcal{A}|$ in (11.2) on the AUC results of Hyper-Kvasir testing set, where PaDiM [43] is used as the anomaly detector.

Methods	AUC	Specificity	Sensitivity	Accuracy
ImageNet	0.943	0.846	0.929	0.898
SimCLR [31]	0.945	0.794	0.942	0.914
Rot-Net [77]	0.938	0.856	0.905	0.905
CSI [214]	0.946	0.952	0.914	0.933
DROC [208]	0.931	0.954	0.881	0.914
SupCon [?]	0.946	0.912	0.953	0.928
CCD [221]	0.978	0.923	0.961	0.967
PMSACL	0.996	0.966	0.981	0.983

Table 5.9: Ablation studies with different self-supervised pre-training approaches on Hyper-Kvasir testing set. PaDiM [43] is used as the anomaly detector. Best results are highlighted.

rotation is the second best data augmentation approach with an AUC result that is 5.1% worse than MedMix. Other approaches do not work well with the appearance characteristics of X-ray images, yielding significantly worse results than our MedMix on Covid-X. These results suggest that the use of MedMix as the strong augmentation yields the best AUC results on different medical image benchmarks.

MedMix Augmentations

In Fig. 5.8, we explore the influence of the number of MedMix augmentation distributions (i.e., $|\mathcal{A}|$) on the AUC results of Hyper-Kvasir, based on our self-supervised PMSACL pre-training and PaDiM anomaly detector. Our model achieves the best performance when $|\mathcal{A}| = 4$ strong augmentation distributions, when it reaches around 98% to 99% AUC. The AUC declines when $|\mathcal{A}| > 5$ or $|\mathcal{A}| < 3$. The performance deterioration when $|\mathcal{A}| < 3$ is due to an insufficient number of pseudo abnormal train-

ing regions from the strong augmentation distributions. When the number of strong augmentation distributions increases to $|\mathcal{A}| > 5$, the pseudo abnormalities may hide most of the normal image regions, causing the model to become over-confident when classifying the pseudo abnormal regions.

Other Self-supervised Methods

In Table 5.9, we show the results of different pre-training approaches with PaDiM as anomaly detector, on Hyper-Kvasir testing set. It can be observed that our PMSACL approach surpasses the previous SOTA CCD pre-training [221] by 2.2% AUC. Other pre-training methods proposed in computer vision (e.g., ImageNet pre-training, SimCLR [31], Rot-Net [77]) achieve worse results than CCD and PMSACL. An interesting point in this comparison is the relatively poor result from ImageNet pre-training, suggesting that it may not generalise well for anomaly detection in medical images. Finally, our PMSACL achieves better results than previous SOTA UAD SSL approaches CSI [214] and DROC [208] by about 4% to 5% AUC, indicating the effectiveness of our new contrastive loss. We also compare the SOTA supervised contrastive learning SupCon [?], which re-formulates the contrastive loss as a supervised task. To validate the effectiveness of our proposed PMSACL contrastive loss, we adapted SupCon to our pseudo multi-class pre-training paradigm for performance comparison. The anomaly detection performance of our PMSACL significantly surpasses SupCon. We argue that such performance improvement is due to the fact that SupCon does not contrast the samples from same classes, missing the chance of learning fine-grained normality features between those samples.

5.5 Conclusion

In this chapter, we proposed a new self-supervised pre-training approach, namely PMSACL, for UAD methods applied to MIA problems. PMSACL is based on a new contrastive learning optimisation to learn multiple classes of normal and pseudo abnormal images, formed with the proposed MedMix data augmentation that simulates medical abnormalities. After pre-training a UAD model using our PMSACL, we fine-tune it with two SOTA anomaly detecting approaches. Experimental results indicate that our PMSACL pre-training can effectively improve the performance of anomaly detection and segmentation on several medical datasets for both anomaly detectors. In the future, we plan to design a new anomaly detector that suits better the characteristics of our self-supervised PMSACL pre-training.

Statement of Authorship

Title of Paper	Deep One-Class Classification via Interpolated Gaussian Descriptor
Publication Status	<input checked="" type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	Published in Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI) 2022 Oral.

Principal Author

Name of Principal Author (Candidate)	Yu Tian		
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper (Equal contribution with Yuanhong Chen).		
Overall percentage (%)	40		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature	_____	Date	09/03/2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Yuanhong Chen		
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper.		
Signature	_____	Date	09/03/2022

Name of Co-Author	Guansong Pang		
Contribution to the Paper	Discussion and writing the revision.		
Signature	_____	Date	09/03/2022

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Please cut and paste additional co-author panels here as required.

Chapter 6

Deep One-Class Classification via Interpolated Gaussian Descriptor

Abstract

One-class classification (OCC) aims to learn an effective data description to enclose all normal training samples and detect anomalies based on the deviation from the data description. Current state-of-the-art OCC models learn a compact normality description by hyper-sphere minimisation, but they often suffer from overfitting the training data, especially when the training set is small or contaminated with anomalous samples. To address this issue, we introduce the interpolated Gaussian descriptor (IGD) method, a novel OCC model that learns a one-class Gaussian anomaly classifier trained with adversarially interpolated training samples. The Gaussian anomaly classifier differentiates the training samples based on their distance to the Gaussian centre and the standard deviation of these distances, offering the model a discriminability w.r.t. the given samples during training. The adversarial interpolation is enforced to consistently learn a smooth Gaussian descriptor, even when the training data is small or contaminated with anomalous samples. This enables our model to learn the data description based on the representative normal samples rather than fringe or anomalous samples, resulting in significantly improved normality description. In extensive experiments on diverse popular benchmarks, including MNIST, Fashion MNIST, CIFAR10, MVTec AD and two medical datasets, IGD achieves better detection accuracy than current state-of-the-art models. IGD also shows better robustness in problems with small or contaminated training sets.

6.1 Introduction

Anomaly detection and segmentation are critical tasks in many real-world applications, such as the identification of defects on industry objects [13] or abnormalities from

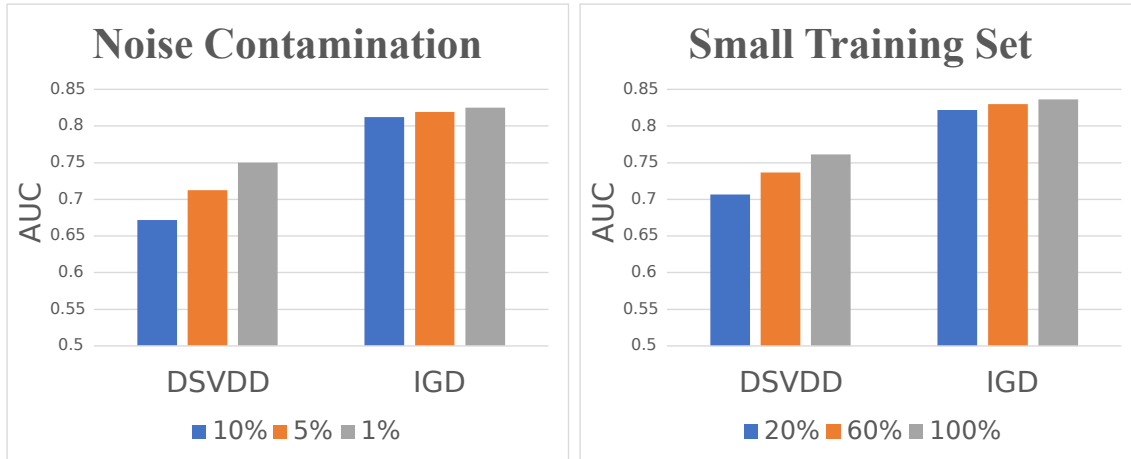


Figure 6.1: Mean testing AUC of DSVDD [191], and our proposed IGD trained with the CIFAR10 training set contaminated with 1%, 5% and 10% of anomalous samples (left), and small training sets, consisting of 20%, 60%, and 100% of the CIFAR10 training set (right).

medical images [199, 200]. Given that most of the training sets available for this task contain only normal images, existing methods are typically formulated as one-class classifiers (OCC) [191, 228]. OCCs aim to first learn a data description of normal samples in the training set and then use a criterion (e.g., distance to the one-class centre [191]) to detect and localise anomalies in test samples.

State-of-the-art (SOTA) OCC models are trained by minimising the radius of a hyper-sphere to enclose all training samples in the representation space [177, 191, 193]. To avoid catastrophic collapse, where all training samples are projected to a single point in the representation space, these OCC models fix the hyper-sphere centre and remove the bias terms from the model. Even though these SOTA OCC models show accurate anomaly detection results in several benchmarks, they can overfit the training data, particularly when the training set is small or contaminated with anomalous samples, as shown by the results of DSVDD [191] in Fig. 6.1.

In this paper, we introduce the interpolated Gaussian descriptor (IGD) method to address the overfitting issue presented in SOTA OCC models. IGD is based on a one-class Gaussian anomaly classifier modelled with adversarially interpolated training samples. The classifier is trained to build a normality description to discriminate training samples based on their distance to the Gaussian centre and the standard deviation of these distances. The smoothness of the normality description is enforced by the adversarial interpolation of the training samples that constrains the training of IGD to be based on representative normal samples rather than fringe or anomalous samples. This allows the normality description of IGD to be more robust than the

SOTA OCC models, particularly when the training set is small or contaminated with anomalous samples, as shown in Fig. 6.1 and t-SNE results in appendix.

In summary, our paper makes the following contributions:

- One novel OCC model that targets the learning of an effective normality description based on representative normal samples rather than fringe or anomalous samples, resulting in an improved anomaly classifier, compared with the SOTA;
- One new OCC optimisation approach based on a theoretically sound derivation of the expectation-maximisation (EM) algorithm that optimises a Gaussian anomaly classifier constrained by adversarially interpolated training samples and multi-scale structural and non-structural image reconstruction to enforce a smooth normality description; and
- One new OCC benchmark to assess the robustness of anomaly detectors to training sets that are small or contaminated with anomalous samples.

Extensive empirical results on six popular anomaly detection benchmarks for semantic anomaly detection, industrial defect detection, and malignant lesion detection show that our model IGD can generalise well across these diverse application domains and perform consistently better than current SOTA detectors. We also show that IGD is more robust than current OCC approaches when dealing with small and contaminated training sets.

6.2 Related Work

Unsupervised anomaly detection (UAD) is generally solved with OCCs [10, 11, 43, 77, 120, 176, 191, 197, 216, 217, 218, 221, 236, 237, 260]. A representative OCC model is DSVDD [191], which forces normal image features to be inside a hyper-sphere with a pre-defined centre and a radius that is minimised to include all training images. Then, test images that fall inside the hyper-sphere are classified as normal, and the ones outside are anomalous. Although powerful, the hard boundary of SVDD can cause the model to overfit the training data – this problem was tackled with a soft-boundary SVDD [191], but it can still overfit given that it lacks enough generalisation constraints. OCC methods can also rely on generative models, such as generative adversarial network (GAN) or Auto-encoder (AE). In [176], a GAN is trained to produce normal samples, and its discriminator is used to detect anomalies, but the complex training process of GANs represents a disadvantage of this approach. An AE [56, 100, 157, 194, 195, 228] is trained to reconstruct normal data, and the anomaly score is defined as the reconstruction error between the input and reconstructed images. AE approaches depend on the MSE reconstruction loss, which does not work well for structural anomalies. Alternatively, single-scale SSIM loss [15] tends to work well for structural anomalies

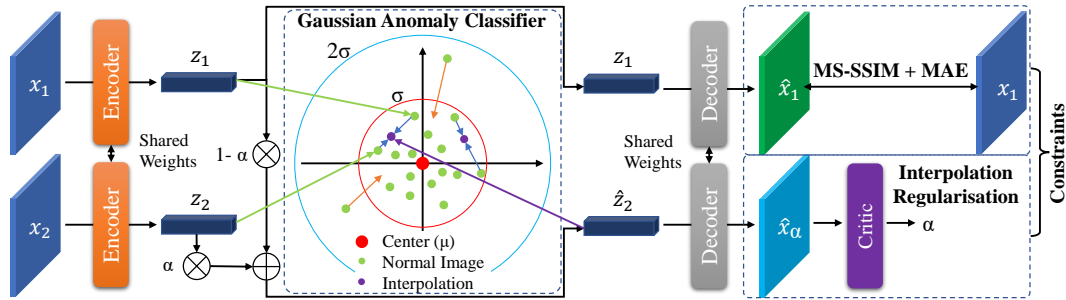


Figure 6.2: Our IGD consists of an encoder that transforms image \mathbf{x} into representation \mathbf{z} , a decoder to reconstruct the image (trained with MS-SSIM and MAE losses), a Gaussian anomaly classifier trained to push the normal image representation close to the centre of the estimated normal image distribution (denoted by a Gaussian with mean μ and standard deviation σ), and a critic module that constrains the likelihood maximisation by predicting the interpolation coefficient α that produces a convex combination of training sample representations. Note that critic is a module similar to a GAN discriminator.

of a specific size, but it may work poorly for non-structural anomalies and structural anomalies outside that specific size. A more detailed review of these methods can be found in [165].

An important aspect of current UAD approaches is their dependence on pre-trained models to produce SOTA results. UAD models can be pre-trained on ImageNet [11, 228] or self-supervised tasks [10, 77]. To allow a fair comparison with current UAD methods, we pre-train IGD with self-supervision and ImageNet.

Unsupervised anomaly localisation targets the segmentation of anomalous image pixels or patches, containing, for example, lesions in medical images [122], defects in industry images [11, 13], or road anomalies in traffic images [173, 219]. The main idea explored is based on extending the image based OCC to a pixel-based OCC, where testing produces a pixel-wise anomaly score map [8, 15]. In general, methods that can localise anomalies [11, 228] are tuned to particular anomaly sizes and structure, which can cause then to miss anomalies outside that range of sizes and structure. To avoid this issue, we design IGD to detect multi-scale structural and non-structural anomalies to improve the anomaly localisation accuracy.

6.3 Method

We denote the training set containing only normal samples by $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{W \times H \times 3}$ represents an RGB image of width W and height H and sampled from

the distribution of normal images as in $\mathbf{x} \sim \mathcal{P}_{\mathcal{X}}$. The testing set contains normal and anomalous images, where anomalous images can have segmentation map annotations. This testing set is defined by $\mathcal{T} = \{(\mathbf{x}_i, y_i, \mathbf{b}_i^{(y_i)})\}_{i=1}^{|\mathcal{T}|}$, where $y_i \in \mathcal{Y} = \{0, 1\}$ (0 denotes a normal and 1 denotes an anomalous image), the segmentation map with the anomaly is denoted by $\mathbf{b}_i^{(y_i)} \in \{0, 1\}^{W \times H}$ (i.e., a pixel-wise anomaly map for image \mathbf{x}_i) if $y_i = 1$, and $\mathbf{b}_i^{(y_i)} = 0^{W \times H}$ if $y_i = 0$.

6.3.1 Interpolated Gaussian Descriptor (IGD)

As depicted in Fig. 6.2, the IGD model is represented by the general classifier $p_{\theta}(y = 0|\mathbf{x}, \mathcal{P}_{\mathcal{X}})$ that consists of an encoder $\mathbf{z} = f_{\psi}(\mathbf{x})$ that transforms a training sample from the image space \mathcal{X} to a representation space $\mathcal{Z} \in \mathbb{R}^Z$, a Gaussian anomaly classifier $p_{\theta}(y = 0|\omega, \mathbf{x}) \in [0, 1]$ that takes the normal image distribution parameter ω and image \mathbf{x} to estimate the probability that it is normal, a decoder $\hat{\mathbf{x}} = g_{\phi}(\mathbf{z})$ that reconstructs an image from the representation space, and a critic module $\alpha = d_{\eta}(g_{\phi}(\alpha\mathbf{z}_1 + (1 - \alpha)\mathbf{z}_2))$ that predicts the interpolation constraint parameter $\alpha \in [0, 1]$, with $\mathbf{z}_1, \mathbf{z}_2$ obtained from the encoder $f_{\psi}(\cdot)$. The IGD parameter $\theta \in \Theta$ represents all module parameters $\{\psi, \phi, \eta\}$ and is estimated with maximum likelihood estimation (MLE):

$$\theta^* = \arg \max_{\theta} \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}_i \in \mathcal{D}} \log p_{\theta}(y_i = 0|\mathbf{x}_i, \mathcal{P}_{\mathcal{X}}). \quad (6.1)$$

We train the one-class classifier in (6.1) using an EM optimisation [46], where the mean and standard deviation of the normal image distribution are estimated during the E-step, instead of being explicitly optimised [191], reducing the risk of overfitting. To encourage the M-step to learn an effective normality description (such that the optimisation is robust to small and contaminated training sets), we add an adversarial interpolation constraint to enforce linear combinations of normal image representations to belong to the normal distribution. We further increase the robustness of IGD to overfitting by constraining the optimisation of the M-step to enforce accurate image reconstruction from its representation. Below, we provide more details about the training process.

To formulate the EM optimisation, we re-write the log-likelihood in (6.1) as

$$\begin{aligned} \log p_{\theta}(y_i = 0|\mathbf{x}_i, \mathcal{P}_{\mathcal{X}}) \\ = \ell_{ELBO}(q, \theta) + KL[q(\omega)||p_{\theta}(\omega|\mathcal{P}_{\mathcal{X}})]. \end{aligned} \quad (6.2)$$

with $\omega \in \mathcal{W} \subset \mathbb{R}^Z \times \mathbb{R}$ denoting the latent variables (mean and standard deviation) that describe the distribution of normal image representations (defined in more detail below). In (6.2), we remove the conditional dependence of $p_{\theta}(\omega|\mathcal{P}_{\mathcal{X}})$ on $y_i = 0$ and \mathbf{x}_i because ω is a variable for the whole training distribution defined as

$$p_{\theta}(\omega|\mathcal{P}_{\mathcal{X}}) = \delta_a(\|\omega(1) - \mu_x\|_2) \delta_a(\omega(2) - \sigma_x), \quad (6.3)$$

where $\delta_a(b) = \frac{1}{|a|\sqrt{\pi}} \exp -(b/a)^2$ ($a \rightarrow 0$ approximates a Dirac delta function, and $a \rightarrow \infty$ approximates a uniform function), $\mu_x = \mathbb{E}_{\mathbf{x} \sim \mathcal{P}_X}[f_\psi(\mathbf{x})]$ and $\sigma_x^2 = \mathbb{E}_{\mathbf{x} \sim \mathcal{P}_X}[\|f_\psi(\mathbf{x}) - \mu_x\|_2^2]$, with $f_\psi(\cdot)$ representing the encoder; and in (6.2), we also have

$$\begin{aligned} \ell_{ELBO}(q, \theta) = & \\ & \mathbb{E}_{q(\omega)}[\log p_\theta(y_i = 0, \omega | \mathbf{x}_i, \mathcal{P}_X)] - \mathbb{E}_{q(\omega)}[\log q(\omega)], \end{aligned} \quad (6.4)$$

where $KL[\cdot]$ denotes the Kullback-Leibler divergence, and $q(\omega)$ represents the variational distribution that approximates $p_\theta(\omega | \mathcal{P}_X)$, defined in (6.3).

The E-step of the EM optimisation zeroes the KL divergence in (6.2) by setting $q(\omega) = p_{\theta^{old}}(\omega | \mathcal{P}_X)$, where θ^{old} represents the previous EM iteration parameter value. In practice, the E-step sets $\omega(1)$ to μ_x and $\omega(2)$ to σ_x , defined in (6.3). Next, the M-step maximises ℓ_{ELBO} in (6.4), with:

$$\begin{aligned} \theta^* = \arg \max_{\theta} \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}_i \in \mathcal{D}} \left(\mathbb{E}_{q(\omega)}[\log p_\theta(y_i = 0 | \omega, \mathbf{x}_i) \right. \\ \left. + \log p_\theta(\omega | \mathcal{P}_X)] \right), \end{aligned} \quad (6.5)$$

where $\mathbb{E}_{q(\omega)}[\log(q(\omega))]$ is removed from ℓ_{ELBO} because it depends only on the previous iteration parameter θ^{old} , $q(\omega)$ is defined in the E-step above, and the conditional dependence of $p_\theta(y = 0 | \omega, \mathbf{x}_i)$ on \mathcal{P}_X is removed because the information from that distribution is summarised in θ . Therefore, (6.5) has two components: 1) the classification term represented by the Gaussian anomaly classifier $p_\theta(y = 0 | \omega, \mathbf{x}_i) = \exp\left(-\frac{\|f_\psi(\mathbf{x}_i) - \omega(1)\|_2^2}{\omega(2)^2}\right)$, with mean $\omega(1)$ and standard deviation $\omega(2)$; and 2) $p_\theta(\omega | \mathcal{P}_X)$ defined in (6.3), which approximates a uniform distribution to prevent the confirmation bias of the estimated μ_x and σ_x from (6.3). To promote an effective normality description of IGD, we constrain the M-step (6.5) as follows:

$$\begin{aligned} \max_{\theta} \quad & \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}_i \in \mathcal{D}} \mathbb{E}_{q(\omega)}[\log(p_\theta(y = 0 | \omega, \mathbf{x}_i))] \\ \text{s.t.} \quad & \ell_d(\mathbf{x}_i, \theta) = 0, \forall \mathbf{x}_i \in \mathcal{D}, \\ & \ell_{f,g}(\mathbf{x}_i, \theta) = 0, \forall \mathbf{x}_i \in \mathcal{D}, \end{aligned} \quad (6.6)$$

where $\ell_d(\cdot)$ is a constraint, defined in (6.10), to enforce the adversarial linear interpolation of normal image representations to belong to the normal representation distribution, and $\ell_{f,g}(\cdot)$ is a constraint, defined in (6.11), to enforce accurate structural and non-structural multi-scale image reconstruction. Note that the maximisation in (6.6) constrains the optimisation in (6.5), which means that we are maximising a lower bound to the original M-step. Using Lagrange multipliers, the optimisation in (6.6) is

reformulated to minimise the following loss function:

$$\ell(\theta, \omega, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \ell_h(\mathbf{x}_i, \omega, \theta) + \lambda_1 \ell_d(\mathbf{x}_i, \theta) + \lambda_2 \ell_{f,g}(\mathbf{x}_i, \theta), \quad (6.7)$$

where

$$\ell_h(\mathbf{x}, \omega, \theta) = 1 - p_\theta(y = 0 | \omega, \mathbf{x}) = p_\theta(y = 1 | \omega, \mathbf{x}), \quad (6.8)$$

with $p_\theta(y = 0 | \omega, \mathbf{x})$ defined in (6.5), and λ_1, λ_2 denoting the Lagrange multipliers. The interpolation constrain $\ell_d(\cdot)$ in (6.6) and (6.7) regularises the training by linearly interpolating the representations from training images, and estimating the interpolation coefficient with the critic network [16]. This interpolation constrains the normal image distribution denser in the representation space, reducing the likelihood that anomalous representations may land in the same region of the representation space occupied by normal samples. Unlike Mix-up [261], our interpolation constraint is a self-supervised method that does not rely on data augmentation on the input space and does not interpolate training labels, making it more adequate for our problem because it enforces a compact and dense distribution of normal samples to be estimated for the Gaussian anomaly classifier. The critic network is represented by

$$\hat{\alpha} = d_\eta(\hat{\mathbf{x}}_\alpha), \quad (6.9)$$

where $\hat{\mathbf{x}}_\alpha = g_\phi(\alpha \mathbf{z}_1 + (1 - \alpha) \mathbf{z}_2)$ represents the reconstruction of the interpolation of $\mathbf{z}_1 = f_\psi(\mathbf{x}_1)$ and $\mathbf{z}_2 = f_\psi(\mathbf{x}_2)$ (with $\alpha \sim \mathcal{U}(0, 0.5)$, $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$, $\mathbf{x}_1 \neq \mathbf{x}_2$, and \mathcal{U} denoting a uniform distribution) [16], and $g_\phi(\cdot)$ denotes the decoder. The goal of the critic network $d_\eta(\cdot)$ is to predict the interpolation coefficient α . The critic network in (6.9) is similar to the discriminator in GAN [79], and relies on the following adversarial loss to be optimised [16]

$$\ell_d(\mathbf{x}, \theta) = \|d_\eta(\hat{\mathbf{x}}_\alpha) - \alpha\|_2^2 + \|d_\eta(\hat{\mathbf{x}}_\zeta)\|_2^2, \quad (6.10)$$

where $\hat{\mathbf{x}}_\alpha$ is defined in (6.9), and $\hat{\mathbf{x}}_\zeta = \zeta \mathbf{x} + (1 - \zeta) \hat{\mathbf{x}}$, with $\zeta \sim \mathcal{U}(0, 1)$ and $\hat{\mathbf{x}}$ denoting a reconstruction of \mathbf{x} by the auto-encoder. The first term of (6.10) minimises the critic's prediction error for α and the second term regularises the training to ensure that the critic predicts $\hat{\alpha} = 0$ when the original image is interpolated with its own reconstruction in the image space \mathcal{X} .

The image reconstruction constrain $\ell_{f,g}(\cdot)$ in (6.6) and (6.7) is defined as

$$\ell_{f,g}(\mathbf{x}, \theta) = \ell_r(\mathbf{x}, \hat{\mathbf{x}}, \theta) + \lambda_3 \|d_\eta(\hat{\mathbf{x}}_\alpha)\|_2^2, \quad (6.11)$$

where $\hat{\mathbf{x}}$ is a reconstruction of \mathbf{x} by the auto-encoder, with the image reconstruction loss $\ell_r(\cdot)$ to be defined below in (6.12), and λ_3 is a hyperparameter to weight the regularisation term. This regularisation fools the critic to output $\hat{\alpha} = 0$ for interpolated

embeddings, independently of α , following standard adversarial training [79]. In (6.11), we also have

$$\begin{aligned} \ell_r(\mathbf{x}, \hat{\mathbf{x}}, \theta) = \\ \sum_{\omega \in \Omega} \rho |\mathbf{x}(\omega) - \hat{\mathbf{x}}(\omega)| + (1 - \rho) \left(1 - m(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega)) \right), \end{aligned} \quad (6.12)$$

with Ω denoting the image lattice, $\rho \in [0, 1]$, $|\mathbf{x}(\omega) - \hat{\mathbf{x}}(\omega)|$ representing the MAE loss, and $m(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega)) \in [0, 1]$ being the MS-SSIM score [242], with larger values indicating higher similarity between patches $\omega \in \Omega$ of the original and reconstructed images.

The loss in (6.7) is used to train two models. A global model that works on the whole image \mathbf{x} , and a local model that works on image patches $\mathbf{x}^{(L)}(\omega) \in \mathcal{X}^{(L)} \subset \mathbb{R}^{W^{(L)} \times H^{(L)} \times 3}$, with $W^{(L)} < W$ and $H^{(L)} < H$, centred at pixel $\omega \in \Omega$ (Ω is the image lattice). During inference, the results from the global and local models are combined to produce multi-scale anomaly detection and localisation.

6.3.2 Theoretical Guarantees

IGD maximises a constrained $\ell_{ELBO}(q, \theta)$ in (6.6) rather than maximising $p_\theta(y = 0 | \mathbf{x}, \mathcal{P}_\mathcal{X})$ in (6.1). Using Theorem 1 in [46], Lemma A.6.1 demonstrates the correctness of IGD, where an increase to the constrained $\ell_{ELBO}(q, \theta)$ implies an increase to $p_\theta(y = 0 | \mathbf{x}, \mathcal{P}_\mathcal{X})$. Using Theorem 2 in [46], Lemma A.7.1 proves the convergence conditions of IGD.

Lemma 6.3.1. *Assuming that the maximisation of the constrained ℓ_{ELBO} in (6.6) produces θ that makes*

$\mathbb{E}_{q(\omega)}[\log p_\theta(y = 0, \omega | \mathbf{x}, \mathcal{P}_\mathcal{X})] \geq \mathbb{E}_{q(\omega)}[\log p_{\theta^{old}}(y = 0, \omega | \mathbf{x}, \mathcal{P}_\mathcal{X})]$,
we have that $(\log p_\theta(y = 0 | \mathbf{x}, \mathcal{P}_\mathcal{X}) - \log p_{\theta^{old}}(y = 0 | \mathbf{x}, \mathcal{P}_\mathcal{X}))$ is lower bounded by
 $(\mathbb{E}_{q(\omega)}[\log p_\theta(y = 0, \omega | \mathbf{x}, \mathcal{P}_\mathcal{X})] - \mathbb{E}_{q(\omega)}[\log p_{\theta^{old}}(y = 0, \omega | \mathbf{x}, \mathcal{P}_\mathcal{X})]) \geq 0$,
with $q(\omega) = p_{\theta^{old}}(\omega | \mathcal{P}_\mathcal{X})$.

Proof. We follow the proof for Theorem 1 in [46]. From the main paper, we have

$$\begin{aligned} \log p_\theta(y = 0 | \mathbf{x}, \mathcal{P}_\mathcal{X}) = \\ \ell_{ELBO}(q, \theta) + KL[q(\omega) || p_\theta(\omega | \mathcal{P}_\mathcal{X})], \end{aligned} \quad (6.13)$$

where $q(\omega) = p_{\theta^{old}}(\omega | \mathcal{P}_\mathcal{X})$. Subtracting $\log p_\theta(y = 0 | \mathbf{x}, \mathcal{P}_\mathcal{X})$ and $\log p_{\theta^{old}}(y = 0 | \mathbf{x}, \mathcal{P}_\mathcal{X})$, we have

$$\begin{aligned} \log p_\theta(y = 0 | \mathbf{x}) - \log p_{\theta^{old}}(y = 0 | \mathbf{x}) = \\ \ell_{ELBO}(q, \theta) - \ell_{ELBO}(q, \theta^{old}) + \\ KL[q(\omega) || p_\theta(\omega | \mathcal{P}_\mathcal{X})] - KL[q(\omega) || p_{\theta^{old}}(\omega | \mathcal{P}_\mathcal{X})]. \end{aligned} \quad (6.14)$$

Since $KL[q(\omega)||p_\theta(\omega|\mathcal{P}_\mathcal{X})] \geq KL[q(\omega)||p_{\theta^{old}}(\omega|\mathcal{P}_\mathcal{X})]$ and that $\ell_{ELBO}(q, \theta) - \ell_{ELBO}(q, \theta^{old}) = \mathbb{E}_{q(\omega)}[\log p_\theta(y=0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})] - \mathbb{E}_{q(\omega)}[\log p_{\theta^{old}}(y=0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})]$, we conclude that

$$\begin{aligned} \log p_\theta(y=0|\mathbf{x}, \mathcal{P}_\mathcal{X}) - \log p_{\theta^{old}}(y=0|\mathbf{x}, \mathcal{P}_\mathcal{X}) &\geq \\ \mathbb{E}_{q(\omega)}[\log p_\theta(y=0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})] - & \\ \mathbb{E}_{q(\omega)}[\log p_{\theta^{old}}(y=0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})] &\geq 0 \end{aligned} \quad (6.15)$$

because of the assumption in this Lemma. \square

Lemma 6.3.2. *Assume that $\{\theta^{(e)}\}_{e=1}^{+\infty}$ denotes the sequence of trained model parameters from the constrained optimisation of ℓ_{ELBO} in (6.6) such that: 1) the sequence $\{\log p_{\theta^{(e)}}(y=0|\mathbf{x}, \mathcal{P}_\mathcal{X})\}_{e=1}^{+\infty}$ is bounded above, and 2) $(\mathbb{E}_{q(\omega)}[\log p_{\theta^{(e+1)}}(y=0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})] - \mathbb{E}_{q(\omega)}[\log p_{\theta^{(e)}}(y=0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})]) \geq \xi (\theta^{(e+1)} - \theta^{(e)})^\top (\theta^{(e+1)} - \theta^{(e)})$, for $\xi > 0$ and all $e \geq 1$, and $q(\omega) = p_{\theta^{(e)}}(\omega|\mathcal{P}_\mathcal{X})$. Then $\{\theta^{(e)}\}_{e=1}^{+\infty}$ converges to some $\theta^* \in \Theta$.*

Proof. We follow the proof for Theorem 2 in [46]. The sequence $\{\log p_{\theta^{(e)}}(y=0|\mathbf{x}, \mathcal{P}_\mathcal{X})\}_{e=1}^{+\infty}$ is non-decreasing (from Lemma A.6.1) and bounded above (from assumption (1) in Lemma A.7.1), so it converges to $L^* < +\infty$. Hence, using Cauchy criterion [158], for any $\epsilon > 0$, we have $e^{(\epsilon)}$ such that, for $e \geq e^{(\epsilon)}$ and all $r \geq 1$,

$$\begin{aligned} \sum_{j=1}^r (\log p_{\theta^{(e+j)}}(y=0|\mathbf{x}, \mathcal{P}_\mathcal{X}) - \log p_{\theta^{(e+j-1)}}(y=0|\mathbf{x}, \mathcal{P}_\mathcal{X})) &= \\ (\log p_{\theta^{(e+r)}}(y=0|\mathbf{x}, \mathcal{P}_\mathcal{X}) - \log p_{\theta^{(e)}}(y=0|\mathbf{x}, \mathcal{P}_\mathcal{X})) &< \epsilon. \end{aligned} \quad (6.16)$$

From (A.7),

$$\begin{aligned} 0 &\leq \mathbb{E}_{q(\omega)}[\log p_{\theta^{(e+j)}}(y=0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})] - \\ &\mathbb{E}_{q(\omega)}[\log p_{\theta^{(e+j-1)}}(y=0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})] \\ &\leq \log p_{\theta^{(e+j)}}(y=0|\mathbf{x}, \mathcal{P}_\mathcal{X}) - \log p_{\theta^{(e+j-1)}}(y=0|\mathbf{x}, \mathcal{P}_\mathcal{X}) \end{aligned} \quad (6.17)$$

for $j \geq 1$ and $q(\omega) = p_{\theta^{(e+j-1)}}(\omega|\mathcal{P}_\mathcal{X})$. Hence, from (A.8),

$$\begin{aligned} \sum_{j=1}^r (\mathbb{E}_{q(\omega)}[\log p_{\theta^{(e+j)}}(y=0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})] - \\ \mathbb{E}_{q(\omega)}[\log p_{\theta^{(e+j-1)}}(y=0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})]) &< \epsilon, \end{aligned} \quad (6.18)$$

for $e \geq e^{(\epsilon)}$ and all $r \geq 1$. Given assumption (2) in Lemma A.7.1 for $e, e+1, e+2, \dots, e+r-1$, we have from (A.10),

$$\epsilon > \xi \sum_{j=1}^r (\theta^{(e+j)} - \theta^{(e+j-1)})^\top (\theta^{(e+j)} - \theta^{(e+j-1)}), \quad (6.19)$$

so

$$\epsilon > \xi (\theta^{(e+r)} - \theta^{(e)})^\top (\theta^{(e+r)} - \theta^{(e)}), \quad (6.20)$$

which is a requirement to prove the convergence of $\theta^{(e)}$ to some $\theta^* \in \Theta$. \square

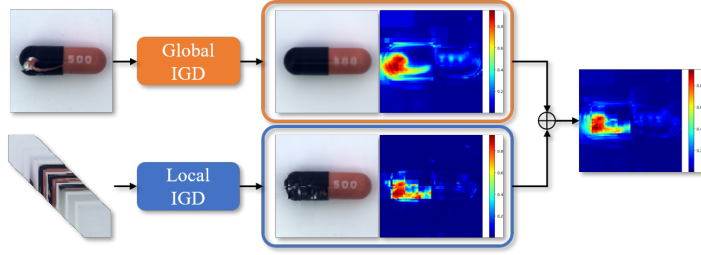


Figure 6.3: Example of the multi-scale structural and non-structural anomaly localisation result for an MVTEC AD [13] image, using both the local and global IGD models. The global model tends to produce smooth results but with some mistakes, while the local model produces jagged results, but without the global mistakes, so by combining the two results, we obtain a smooth and correct anomaly heatmap.

6.3.3 Training and Inference

The global and local IGD models are trained separately (see Fig. A.1), following the EM optimisation, where the E-step estimates the the latent variable ω in (6.3), and the M-step minimises the loss in (6.7) to obtain θ^* .

During inference, **anomaly detection** is performed by combining the global and local IGD anomaly scores for a testing image \mathbf{x} as in:

$$s(\mathbf{x}) = s^{(G)}(\mathbf{x}) + s^{(L)}(\mathbf{x}). \quad (6.21)$$

The global score in (6.21) is defined as

$$s^{(G)}(\mathbf{x}) = \ell_r^{(G)}(\mathbf{x}, \hat{\mathbf{x}}, \theta^*) + \ell_h^{(G)}(\mathbf{x}, \theta^*), \quad (6.22)$$

where $\ell_r^{(G)}(\cdot)$ denotes the reconstruction loss from (6.12) and $\ell_h^{(G)}(\cdot)$ denotes the Gaussian anomaly classification loss from (6.8) (both computed with the global IGD model using the whole images), and $\hat{\mathbf{x}}$ is the reconstruction of \mathbf{x} produced by the auto-encoder. The local score in (6.21) is defined as

$$s^{(L)}(\mathbf{x}) = \max_{\omega \in \Omega} \left(\ell_r^{(L)}(\mathbf{x}^{(L)}(\omega), \hat{\mathbf{x}}^{(L)}(\omega), \theta^*) + \ell_h^{(L)}(\mathbf{x}^{(L)}(\omega), \theta^*) \right), \quad (6.23)$$

where $\ell_r^{(L)}(\cdot)$ and $\ell_h^{(L)}(\cdot)$ are the reconstruction and Gaussian anomaly classification losses computed from the local model, with $\mathbf{x}^{(L)}(\omega)$ denoting an image patch of size $W^{(L)} \times H^{(L)} \times 3$ at pixel $\omega \in \Omega$. The use of max pooling of the local scores in (6.23) facilitates detection of images that contain anomalies covering a small region of the image.

In particular, the MS-SSIM loss uses the MS-SSIM global score, defined as

$$m^{(G)}(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega)) = [l_M(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega))]^{\alpha_M} \times \prod_{m=1}^{m^{(G)}} [c_m(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega))]^{\beta_m} [s_m(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega))]^{\gamma_m}, \quad (6.24)$$

where $\mathbf{x}(\omega)$ denotes an image patch centred at $\omega \in \Omega$ of size $11 \times 11 \times 3$,

$$l_M(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega)) = \frac{2\mu_{\mathbf{x}(\omega)}\mu_{\hat{\mathbf{x}}(\omega)} + C_1}{\mu_{\mathbf{x}(\omega)}^2 + \mu_{\hat{\mathbf{x}}(\omega)}^2 + C_1}, \quad (6.25)$$

$$c_m(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega)) = \frac{2\sigma_{\mathbf{x}(\omega)}\sigma_{\hat{\mathbf{x}}(\omega)} + C_2}{\sigma_{\mathbf{x}(\omega)}^2 + \sigma_{\hat{\mathbf{x}}(\omega)}^2 + C_2}, \quad (6.26)$$

$$s_m(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega)) = \frac{\sigma_{\mathbf{x}(\omega)\hat{\mathbf{x}}(\omega)} + C_3}{\sigma_{\mathbf{x}(\omega)}\sigma_{\hat{\mathbf{x}}(\omega)} + C_3}, \quad (6.27)$$

with C_1, C_2, C_3 representing pre-defined constants, $\mu_{\mathbf{x}(\omega)}$ denoting the mean intensities of $\mathbf{x}(\omega)$, $\sigma_{\mathbf{x}(\omega)}^2$ the variance of $\mathbf{x}(\omega)$, and $\sigma_{\mathbf{x}(\omega)\hat{\mathbf{x}}(\omega)}$ the covariance of $\mathbf{x}(\omega)$ and $\hat{\mathbf{x}}(\omega)$. In (A.1), $m^{(G)} = 5$ denotes the number of scales, $\beta_1 = \gamma_1 = 0.0448$, $\beta_2 = \gamma_2 = 0.2856$, $\beta_3 = \gamma_3 = 0.3001$, $\beta_4 = \gamma_4 = 0.2363$, $\alpha_5 = \beta_5 = \gamma_5 = 0.1333$ [242]. We follow $C_i = (K_i L)^2$ (for $i \in \{1, 2, 3\}$) according to [240] and define $L = 4.7579$ as the pixel range with $K_1 = 0.01$, $K_2 = 0.03$ and $C_3 = C_2/2$.

The local score $m^{(L)}(\mathbf{x}^{(L)}(\omega), \hat{\mathbf{x}}^{(L)}(\omega))$ is defined in the same way as in (A.1), where $\mathbf{x}^{(L)}(\omega)$ is an image patch centred at $\omega \in \Omega$ of size $3 \times 3 \times 3$, $m^{(L)} = 4$ scales with weights $\beta_1 = \gamma_1 = 0.0516$, $\beta_2 = \gamma_2 = 0.3295$, $\beta_3 = \gamma_3 = 0.3463$, $\alpha_4 = \beta_4 = \gamma_4 = 0.2726$ modified based on the original proportion for $m^{(G)} = 5$.

Anomaly localisation is computed for each pixel $\omega \in \Omega$ to produce a local score

$$l(\mathbf{x}(\omega)) = \ell_r^{(G)}(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega), \theta^*) + \ell_r^{(L)}(\mathbf{x}^{(L)}(\omega), \hat{\mathbf{x}}^{(L)}(\omega), \theta^*), \quad (6.28)$$

with

$$\ell_r^{(G)}(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega), \theta^*) = \rho |\mathbf{x}(\omega) - \hat{\mathbf{x}}(\omega)| + (1 - \rho) \left(1 - m^{(G)}(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega)) \right), \quad (6.29)$$

where ρ and $m^{(G)}(\cdot)$ are defined in (6.12) and $\hat{\mathbf{x}}$ is a reconstruction of \mathbf{x} produced by the global IGD model. The $\ell_r^{(L)}(\mathbf{x}^{(L)}(\omega), \hat{\mathbf{x}}^{(L)}(\omega), \theta^*)$ in (6.28) is similarly defined using the local IGD model. Thus, the anomaly localisation final map is a heatmap with high values representing regions that are likely to contain anomalies.

6.4 Experiments

6.4.1 Datasets and Evaluation Metric

Datasets: We use four computer vision and two medical image datasets to evaluate our methods. The computer vision datasets are MNIST [48], Fashion MNIST [247], CIFAR10 [113] and MVTec AD [13]; and the medical image datasets are Hyper-Kvasir [21] and LAG [121]. MNIST, Fashion MNIST and CIFAR10 have been widely used as benchmarks for image anomaly detection, and we follow the same experimental protocol as described in [191]. CIFAR10 contains 60,000 images with 10 classes. MNIST and Fashion MNIST contain 70,000 images with 10 classes of handwritten digits and fashion products, respectively. MVTec AD [13] contains 5,354 high-resolution real-world images of 15 different industry object and textures. The normal class of MVTec AD is formed by 3,629 training and 467 testing images without defects. The anomalous class has more than 70 categories of defects (such as dents, structural fails, contamination, etc.) and contains 1,258 testing images. MVTec AD provides pixel-wise ground truth annotations for all anomalies in the testing images, allowing the evaluation of anomaly detection and localisation. We also tested our method on two publicly available medical datasets: Hyper-Kvasir [21] and LAG [121] for polyp and glaucoma detection, respectively. For Hyper-Kvasir, we have 1,600 normal images without polyps in the training set and 500 in the testing set; and 1,000 abnormal images containing polyps in the testing set. For LAG, we have 2,343 normal images without glaucoma in the training set; and 800 normal images and 1,711 abnormal images with glaucoma for testing.

Evaluation: For anomaly detection, we assess performance with the area under the receiver operating characteristic curve (AUC) and classification accuracy. On MNIST, Fashion MNIST and CIFAR10, we use the same protocol as other methods in Tab. 6.1, where training uses a single class as the normal data, with the nine remaining classes denoting as semantically anomalous samples, and inference relies on a non-augmented test image. We report the mean AUC over the 10 classes for the above three data sets. On MVTec AD [11, 228], we evaluate anomaly detection with mean AUC and accuracy. Follow previous works [218, 221], we evaluate the methods using AUC for the Hyper-Kvasir and LAG. For anomaly localisation, we follow [228] and compute the mean pixel-level AUC between the generated heatmap and the ground truth segmentation map for each anomalous image in the testing set of MVTec AD.

6.4.2 Implementation Details

We implement our framework using Pytorch. The model was trained with Adam optimiser using a learning rate of 0.0001, weight decay of 10^{-6} , batch size of 64 images, 256 epochs for all dataset. We defined the representation space produced by the encoder to have $Z = 128$ dimensions. Following [76], we set $\rho = 0.15$ to balance the contribution

Pretrain	Method	MNIST	CIFAR10	FMNIST
Scratch	DAE [85]	0.8766	0.5358	-
	VAE [109]	0.9696	0.5833	-
	KDE [19]	0.8140	0.6100	-
	OCSVM [201]	0.9510	0.5860	-
	AnoGAN [200]	0.9127	0.6179	-
	DSVDD [191]	0.9480	0.6481	-
	OCGAN [176]	0.9750	0.6566	-
	PixelCNN [224]	0.6180	0.5510	-
	CapsNet _{PP} [125]	0.9770	0.6120	0.7650
	CapsNet _{RE} [125]	0.9250	0.5310	0.6790
	ADGAN [41]	0.9680	0.6340	-
	LSA [1]	0.9750	0.6410	0.8760
	MemAE [56]	0.9751	0.6088	-
	GradCon [114]	0.9730	0.6640	-
	λ -VAE _u [44]	0.9820	0.7170	0.8730
	ULSLM [243]	0.9490	0.7360	-
SCADN [251]	0.9771	0.6690	-	
Ours	0.9869	0.7433	0.9201	
ImageNet	CAVGA-D _u [228]	0.9860	0.7370	0.8850
	Student-Teacher [11]	0.9935	0.8196	-
	Ours	0.9927	0.8368	0.9357
SSL	Rot-Net [77]	-	0.8160	0.9350
	(author?) [10]	-	0.8820	0.9410
	Ours	-	0.9125	0.9441

Table 6.1: **Anomaly detection:** mean AUC testing results on MNIST, CIFAR10 and Fashion MNIST. The results are split into ‘Scratch’ (without any pre-training), pretrained with ‘ImageNet’, and self-supervised learning (‘SSL’). Bold numbers represent the best result (within 0.5%) for each data set, discriminated by Scratch, SSL or ImageNet.

of MAE and MS-SSIM losses in (6.12) and (6.29). We set $\lambda_1 = \lambda_2 = 1$ in (6.7) and $\lambda_3 = 0.1$ in (6.11), based on cross validation experiments. We use Resnet18 and its reverse architecture as the encoder and decoder for both the global and local IGD models. When computing the accuracy of anomaly detection in MVTec AD, the threshold of the anomaly detection score $s(\mathbf{x})$ in (6.21) (to classify an image as anomalous) is set to 0.5 [228]. To enable a fair comparison between our method and previous approaches in the field [10, 11, 77, 228], we pre-train the encoders for the global and local IGD models either with self-supervised learning (SSL) [29] or ImageNet knowledge distillation (KD) [11, 80].

For this SSL pre-training, we use the SGD optimiser with a learning rate of 0.01, weight decay 10^{-1} , batch size of 32, and 2,000 epochs. Once we obtain the pre-trained encoder with SSL, we remove the MLP layer and attach a linear layer to the backbone with fixed parameters. Note that this SSL is trained from scratch. In contrast to the vanilla self-supervised learning [29] suggesting large batch size, we notice that a medium batch size yields significantly better performance for unsupervised anomaly

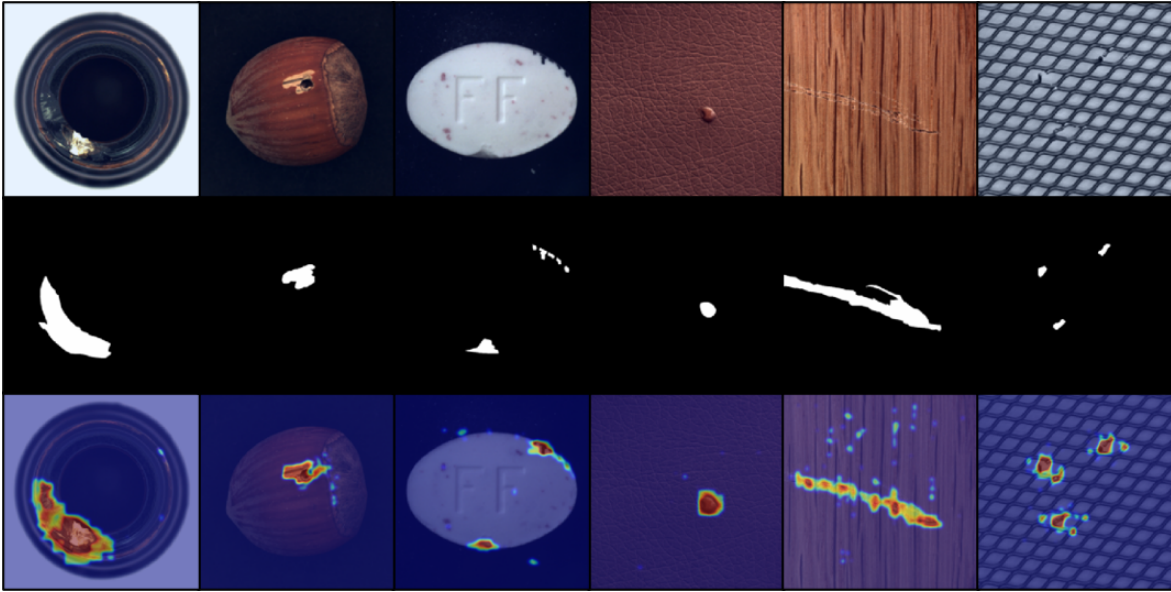


Figure 6.4: Qualitative results of our anomaly localisation results on the MVTec AD (red = high probability of anomaly). Top, middle and bottom rows show the testing images, ground-truth masks and predicted heatmaps, respectively.

detection.

For the ImageNet KD pre-training, we minimise the ℓ_2 norm between the 512-dimensional feature vector output from encoder and an intermediate layer of the ImageNet pre-trained ResNet18 with the same 512-dimensional features. For this ImageNet KD pre-training, we use the Adam optimiser with a learning rate of 0.0001, weight decay 10^{-5} , batch size of 64, and 50,000 iterations. Once we obtain the pre-trained encoder of KD, we fix the network parameters and attach a linear layer to reduce the dimensionality of the feature space to 128.

6.4.3 Experiments on MNIST, Fashion MNIST and CIFAR10

Table 6.1 compares the unsupervised anomaly detection mean AUC testing results between our method and the current SOTA on MNIST, Fashion MNIST and CIFAR10. The rows labelled as ‘Scratch’ show results of models that were not pre-trained, and the ones with ‘SSL’ display results from models using self-supervised learning method [10, 77]. The ones with ‘ImageNet’ show results from models that use ImageNet KD pre-training [11, 228]. Our proposed IGD outperforms current SOTA methods for the majority of pre-training methods on all three datasets.

Metric	Method	Mean
Accuracy	AVID [195]	0.730
	AE _{SSIM} [15]	0.630
	DAE [85]	0.710
	AnoGAN [200]	0.550
	λ -VAE _u [44]	0.770
	LSA [1]	0.730
	CAVGA-D _u [228]	0.780
	CAVGA-R _u [228]	0.820
	Ours - ImageNet	0.840
	Ours - SSL	0.850
AUC	AnoGAN [200]	0.503
	GANomaly [3]	0.782
	Skip-GANomaly [4]	0.805
	SCADN [251]	0.818
	U-Net [189]	0.819
	DAGAN [215]	0.873
	Ours - ImageNet	0.926
	Ours - SSL	0.934

Table 6.2: **Anomaly detection**: mean testing accuracy and AUC on MVTec AD produced by the SOTA and our IGD.

6.4.4 Experiments on MVTec AD

We report the results, based on SSL and ImageNet KD pre-trained models, for both anomaly detection (Tab. 6.2) and localisation (Tab. 6.3) on MVTec AD, which contains real-world images of industry objects and textures containing different types of anomalies. Following [228] the score threshold is set to 0.5 for calculating the mean accuracy of anomaly detection. For anomaly detection, our method produces the best accuracy (at least 2% better than previous SOTA) and AUC (at least 5% better than previous SOTA) results independently of the pre-training technique. For anomaly localisation, we compare our method and the SOTA using the mean pixel-level AUC of all anomalous images in the testing set of MVTec AD. Notice that our method with ImageNet and SSL pre-training are better than the previous SOTA CAVGA-R_u [228] by 2% and 4%, respectively. Fig. 6.4 shows anomaly localisation results on MVTec AD images, where red regions in the heatmap indicate higher anomaly probability. From this results, we can see that our approach can localise anomalous regions of different sizes and structures from different object categories.

Method	MVTec AD
DAE [85]	0.82
AE _{SSIM} [15]	0.87
AVID [195]	0.78
SCADN [251]	0.75
LSA [1]	0.79
λ -VAE _u [44]	0.86
AnoGAN [200]	0.74
ADVAE [136]	0.86
CAVGA-D _u [228]	0.85
CAVGA-R _u [228]	0.89
Ours - ImageNet	0.91
Ours - SSL	0.93

Table 6.3: **Anomaly localisation:** mean pixel-level AUC testing results on the anomalous images of MVTEC AD.

6.4.5 Experiments on Medical Datasets

To show that our method can generalise to other domains, we evaluate our approach on two public medical datasets - Hyper-Kvasir for polyp detection and LAG for glaucoma detection. As shown in Tab. 6.4, our SSL and ImageNet based results achieve the best AUC results on both datasets. Our methods surpass the recent proposed CAVGA-R_u [228] on both datasets by a minimum 0.9% and maximum 3.8%. Also, our model performs better compared to the anomaly detector specifically designed for medical data, such as f-anogan [199] and ADGAN [139].

The abnormalities in medical data (i.e., colon polyps, glaucoma) are significantly different than the popular image benchmarks and MVTEC AD in terms of appearance and structural anomalies, suggesting that our model works in disparate domains.

6.4.6 Visualisation of the Distribution of Testing Samples

Figure A.2 shows the distribution of testing samples in the representation space, using the t-SNE visualisation, for DSVDD [191], Gaussian anomaly classifier (GAC), and our IGD. Notice that the normal samples seem to be more compactly represented with fewer anomalous samples appearing inside the normal cluster. This suggests that IGD has a superior normality description, compared with DSVDD and GAC.

6.4.7 Ablation Study

To investigate the effectiveness of each component of our method, we show the mean AUC results of our method with different proposed variants in Tab. 11.5. Note that all results are based on the initialisation of knowledge distillation from ImageNet.

Methods	Hyper-Kvasir	LAG
DAE [60]	0.705	0.651
CAM [267]	-	0.663
GBP [210]	-	0.787
SmoothGrad [206]	-	0.795
OCGAN [174]	0.813	0.737
F-anoGAN [199]	0.907	0.778
ADGAN [139]	0.913	0.752
CAVGA- R_u [228]	0.928	0.819
Ours - ImageNet	0.931	0.838
Ours - SSL	0.937	0.857

Table 6.4: **Anomaly detection:** AUC testing results on two medical datasets: Hyper-Kvasir and LAG.

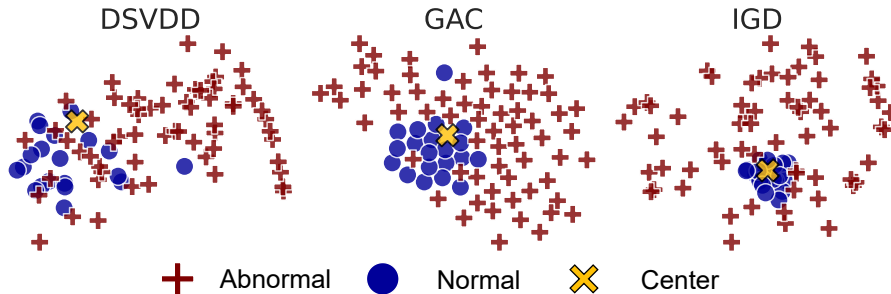


Figure 6.5: t-sne visualisation from MVTec (class bottle).

For standard anomaly detection settings (AUC - Full), each proposed component of our IGD improves performance by a minimum 1.7% and maximum 11.6% mean AUC. Tab. 11.5 also shows the effectiveness of each component when trained with small (20% of full training data) or anomaly contaminated (10% of contamination rate) training sets, where our proposed Gaussian anomaly classifier (GAC) significantly improves over the REC (i.e., MS-SSIM+MAE losses) baseline by 13% and 10.4% mean AUC. The proposed adversarial interpolation regularisation (INTER) further improves the AUC by 3.7% and 3.1%.

6.4.8 Experiments on Small/Contaminated Training Sets

To show the improved robustness of our approach to small training sets on CIFAR10 and MVTec, we compare the performance of DSVDD, DSVDD+REC (i.e., DSVDD combined with our reconstruction loss), and our proposed IGD, using less normal data in the training sets in Tab. 6.6. In particular, we randomly sub-sample 20%, 60%, and 100% of the original training sets of CIFAR10 and MVTec AD, to form a smaller

MSE	REC	GAC	INTER	AUC - Full	AUC - ST	AUC - AC
✓				0.615	0.552	0.565
	✓			0.731	0.655	0.677
	✓	✓		0.819	0.785	0.781
	✓	✓	✓	0.836	0.822	0.812

Table 6.5: Ablation study of our method on CIFAR10 using anomaly detection mean testing AUC w.r.t standard OCC setup (AUC - Full), small training set containing 20% of training data (AUC - ST), and anomaly contaminated training set with 10% contamination (i.e., 10% of the anomalous samples are removed from the testing set and inserted into the training set) (AUC - AC). MSE denotes the baseline deep autoencoder with MSE loss, REC denotes the baseline deep autoencoder with MS-SSIM + MAE losses, GAC denotes our proposed Gaussian anomaly classifier, INTER represents our interpolation regularisation. The encoder of all above methods are initialised based on the knowledge distillation from ImageNet.

Dataset	Train Size	DSVDD	DSVDD+REC	IGD (Ours)
CIFAR10	20%	0.7064	0.7462	0.8219
	60%	0.7367	0.7807	0.8298
	100%	0.7612	0.7950	0.8365
MVTec	20%	0.7994	0.7291	0.9043
	60%	0.8467	0.7737	0.9246
	100%	0.8579	0.7826	0.9260

Table 6.6: Mean testing AUCs on CIFAR10 and MVTec with small training sets, where REC=MS-SSIM+MAE losses.

training set. The results indicate that IGD achieves comparable performance under significantly less training data, while the performance of DSVDD and DSVDD+REC deteriorate dramatically when the number of training samples decreases. This result shows that IGD has better robustness than DSVDD and DSVDD+REC to small training sets.

To show the improved robustness of our approach contaminated training sets, in Tab. 6.7, we compare the performance of DSVDD, DSVDD+REC, and our IGD, using training sets corrupted with anomalous samples (this contamination facilitates overfitting). In particular, we re-organise the original training and test data of CIFAR10 and MVTec AD by randomly sampling 1%, 5% and 10% of anomalies from the test data to inject into the training data. With different rates of anomaly contamination, the maximum fluctuation of our IGD is 1.3% on CIFAR10 and 0.44% on MVTec AD.

Dataset	Noise Ratio	DSVDD	DSVDD+REC	IGD (Ours)
CIFAR10	1%	0.7502	0.7694	0.8252
	5%	0.7124	0.7448	0.8193
	10%	0.6717	0.7073	0.8122
MVTec	1%	0.8523	0.7873	0.9363
	5%	0.8391	0.7733	0.9319
	10%	0.8175	0.7687	0.9363

Table 6.7: Mean testing AUCs on CIFAR10 and MVTec with different contamination noise rates. REC defined in Tab. 6.6.

While the competing method DSVDD shows a much larger maximum fluctuation of 7.8% and 3.5% mean AUC, on CIFAR10 and MVTec AD, respectively. The results show the substantially better robustness of IGD over DSVDD and DSVDD+REC for the anomaly-contaminated training data.

6.5 Discussion

We do not compare some of the SOTA works [185, 208, 214] in Table 6.1, 6.2, and 6.3 due to unfair comparison. In particular, the comparison with PANDA [185] is not fair because it uses a WideResNet50 \times 2 for MVTec and ResNet152 for CIFAR, both being much larger backbones than our ResNet18. Regarding CSI [214], it has much slower inference (because of the 40 \times data augmentation of test images) and more complex training that needs a coreset and large batch size of 512 for pre-training, which challenges its use for problems with small training sets or high-resolution images. For both CSI and DROC [208], their gains are mostly from the SSL pre-training. To show that point for CSI, we use our training approach to fine-tune a pre-trained CSI model and obtain 94.6% AUC on CIFAR10, which is higher than CSI (94.3% AUC). Also, for the vanilla SSL pre-training reported in DROC paper, their performance reduces from 92.5% to 89.0% AUC on CIFAR10, and from 86.5% to 80.2% AUC on MVTec. Note that all above results are collected from their published papers unless stated otherwise.

Furthermore, on MVTec, our approach obtains (93.4% AUC), which is much better than CSI (63.6% AUC from Tab.2 of [186]) and PANDA (86.5%). For anomaly localisation on MVTec, our 93% AUC is better than DROC (90%) and worse than PANDA (96%). On high-resolution image datasets (e.g., Hyper-Kvasir), our approach (93.7% AUC) is better than CSI (trained by us) that reaches 91.6% AUC. Other important results shown by our paper, but missed by CSI, PANDA and DROC, are the ones

with small training sets and contaminated training sets, which are new and important benchmarks for real-world industrial applications and early detection of medical diseases.


6.6 Conclusion

In this paper, we presented a new OCC model, called interpolated Gaussian descriptor (IGD), to perform unsupervised anomaly detection and segmentation. IGD learns a one-class Gaussian anomaly classifier trained with adversarially interpolated training samples to enable an effective normality description based on representative normal samples rather than fringe or anomalous samples. The optimisation of IGD is formulated as an EM algorithm, which we show to be theoretically correct and to converge to a stationary solution under certain conditions. To our knowledge, IGD is the first method that is able to achieve the best performance across diverse application datasets, including MNIST, CIFAR10, Fashion MNIST, MVTec AD, and two large scale medical datasets, in terms of anomaly detection and localisation. We also show that IGD is more robust than DSVDD and an image-reconstruction constrained DSVDD in problems with small or contaminated training sets. We plan to study the use of Gaussian anomaly classifier in the pixel-wise localisation of anomalies and to investigate new self-supervised learning approaches specifically designed for anomaly detection.

Statement of Authorship

Title of Paper	Unsupervised Anomaly Detection in Medical Images with a Memory-augmented Multi-level Cross-attention Masked Autoencoder
Publication Status	<input type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input checked="" type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	Submitted to MICCAI 2022


Principal Author


Name of Principal Author (Candidate)	Yu Tian
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper.
Overall percentage (%)	90
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	 <div style="display: inline-block; border-bottom: 1px solid black; width: 100px; margin-left: 10px;"></div>
Date	09/03/2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Guansong Pang
Contribution to the Paper	Discussion and writing the revision.
Signature	 <div style="display: inline-block; border-bottom: 1px solid black; width: 100px; margin-left: 10px;"></div>
Date	09/03/2022

Name of Co-Author	Yuyuan Liu
Contribution to the Paper	Discussion and writing the revision.
Signature	 <div style="display: inline-block; border-bottom: 1px solid black; width: 100px; margin-left: 10px;"></div>
Date	09/03/2022

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Chong Wang		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Yuanhong Chen		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Fengbei Liu		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Rajvinder Singh		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Johan W Verjans		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Please cut and paste additional co-author panels here as required.

Chapter 7

Unsupervised Anomaly Detection in Medical Images with a Memory-augmented Multi-level Cross-attention Masked Autoencoder

Abstract

Unsupervised anomaly detection (UAD) aims to find anomalous images by optimising a detector using a training set that contains only normal images. UAD approaches can be based on reconstruction methods, self-supervised approaches, and Imagenet pre-trained models. Reconstruction methods, which detect anomalies from image reconstruction errors, are advantageous because they do not rely on the design of problem-specific pretext tasks needed by self-supervised approaches, and on the unreliable translation of models pre-trained from non-medical datasets. However, reconstruction methods may fail because they can have low reconstruction errors even for anomalous images. In this chapter, we introduce a new reconstruction-based UAD approach that addresses this low-reconstruction error issue for anomalous images. Our UAD approach, the memory-augmented multi-level cross-attentional masked autoencoder (MemMC-MAE), is a transformer-based approach, consisting of a novel memory-augmented self-attention operator for the encoder and a new multi-level cross-attention operator for the decoder. MemMC-MAE masks large parts of the input image during its reconstruction, reducing the risk that it will produce low reconstruction errors because anomalies are likely to be masked and cannot be reconstructed. However, when the anomaly is not masked, then the normal patterns stored in the encoder’s memory combined with the decoder’s multi-level cross-attention will constrain the accurate reconstruc-

tion of the anomaly. We show that our method achieves SOTA anomaly detection and localisation on colonoscopy and Covid-19 Chest X-ray datasets.

7.1 Introduction

Detecting and localising anomalous findings in medical images (e.g., polyps, malignant tissues, etc.) are of vital importance [9, 67, 70, 131, 134, 135, 146, 216?]. Systems that can tackle these tasks are often formulated with a classifier trained with large-scale datasets annotated by experts. Obtaining such annotation is often challenging in real-world clinical datasets because the amount of normal images from healthy patients tend to overwhelm the amount of anomalous images. Hence, to alleviate the challenges of collecting anomalous images and learning from class-imbalanced training sets, the field has developed unsupervised anomaly detection (UAD) models [33, 221] that are trained exclusively with normal images. Such UAD strategy benefits from the straightforward acquisition of training sets containing only normal images and the potential generalisability to unseen anomalies without collecting all possible anomalous sub-classes.

Current UAD methods learn a one-class classifier (OCC) using only normal/healthy training data, and detect anomalous/disease samples using the learned OCC [33, 56, 120, 139, 167, 199, 204, 219, 229]. UAD methods can be divided into: 1) reconstruction methods, 2) self-supervised approaches, and 3) Imagenet pre-trained models. Reconstruction methods [33, 56, 139, 199, 229] are trained to accurately reconstruct normal images, exploring the assumption that the lack of anomalous images in the training set will prevent a low error reconstruction of an test image that contains an anomaly. However, this assumption is not met in general because reconstruction methods are indeed able to successfully reconstruct anomalous images, particularly when the anomaly is subtle. Self-supervised approaches [208, 221?] train models using contrastive learning, where pretext tasks must be designed to emulate normal and anomalous image changes for each new anomaly detection problem. Imagenet pre-trained models [185?] produce features to be used by OCC, but the translation of these models into medical image problems is not straightforward. Reconstruction methods are able to circumvent the aforementioned challenges posed by self-supervised and Imagenet pre-trained UAD methods, and they can be trained with a relatively small amount of normal samples. However, their viability depends on an acceptable mitigation of the potentially low reconstruction error of anomalous test images.

In this chapter, we introduce a new UAD reconstruction method, the Memory-augmented Multi-level Cross-attention Masked Autoencoder (MemMC-MAE), designed to address the low reconstruction error of anomalous test images. MemMC-MAE is a transformer-based approach based on masked autoencoder (MAE) [89] with of a novel memory-augmented self-attention encoder and a new multi-level cross-attention de-

coder. MemMC-MAE masks large parts of the input image during its reconstruction, and given that the likelihood of masking out an anomalous region is large, then it is unlikely that it will accurately reconstruct that anomalous region. However, there is still the risk that the anomaly is not masked out, so in this case, the normal patterns stored in the encoder’s memory combined with the correlation of multiple normal patterns in the image, utilised by the decoder’s multi-level cross-attention can explicitly constrain the accurate anomaly reconstruction to produce high reconstruction error (high anomaly score). The encoder’s memory is also designed to address the MAE’s long-range ‘forgetting’ issue [151], which can be harmful for UAD due to the poor reconstruction based on forgotten normality patterns and ‘unwanted’ generalisability to subtle anomalies during testing. Our contributions are summarised as:

- To the best of our knowledge, this is the first UAD method based on MAE [89];
- A new memory-augmented self-attention operator for our MAE transformer encoder to explicitly encode and memorise the normality patterns; and
- A novel decoder architecture that uses the learned multi-level memory-augmented encoder information as prior features to a cross-attention operator.

Our method achieves better anomaly detection and localisation accuracy than most competing approaches on the UAD benchmarks using the public Hyper-Kvasir colonoscopy dataset [21] and Covid-X Chest X-ray (CXR) dataset [235].

7.2 Method

7.2.1 Memory-augmented Multi-level Cross-attention Masked Autoencoder (MemMC-MAE)

Our MemMC-MAE, depicted in Fig. 7.1, is based on the masked autoencoder (MAE) [89] that was recently developed for the pre-training of models to be used in downstream computer vision tasks. MAE has an asymmetric architecture, with a encoder that takes a small subset of the input image patches and a smaller/lighter decoder that reconstructs the original image based on the input tokens from visible patches and dummy tokens from masked patches.

Our MemMC-MAE is trained with a normal image training set, denoted by $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{H \times W \times R}$ (H : height, W : width, R : number of colour channels). Our method first divides the input image \mathbf{x} into non-overlapping patches $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^{|\mathcal{P}|}$, where $\mathbf{p} \in \mathbb{R}^{\hat{H} \times \hat{W} \times R}$, with $\hat{H} \ll H$ and $\hat{W} \ll W$. We then randomly mask out 75% of the $|\mathcal{P}|$ patches, and the remaining visible patches $\mathcal{P}^{(v)} = \{\mathbf{p}_v\}_{v=1}^{|\mathcal{P}^{(v)}|}$ (with $|\mathcal{P}^{(v)}| = 0.25 \times |\mathcal{P}|$) are used by the MemMC-MAE to encode the normality

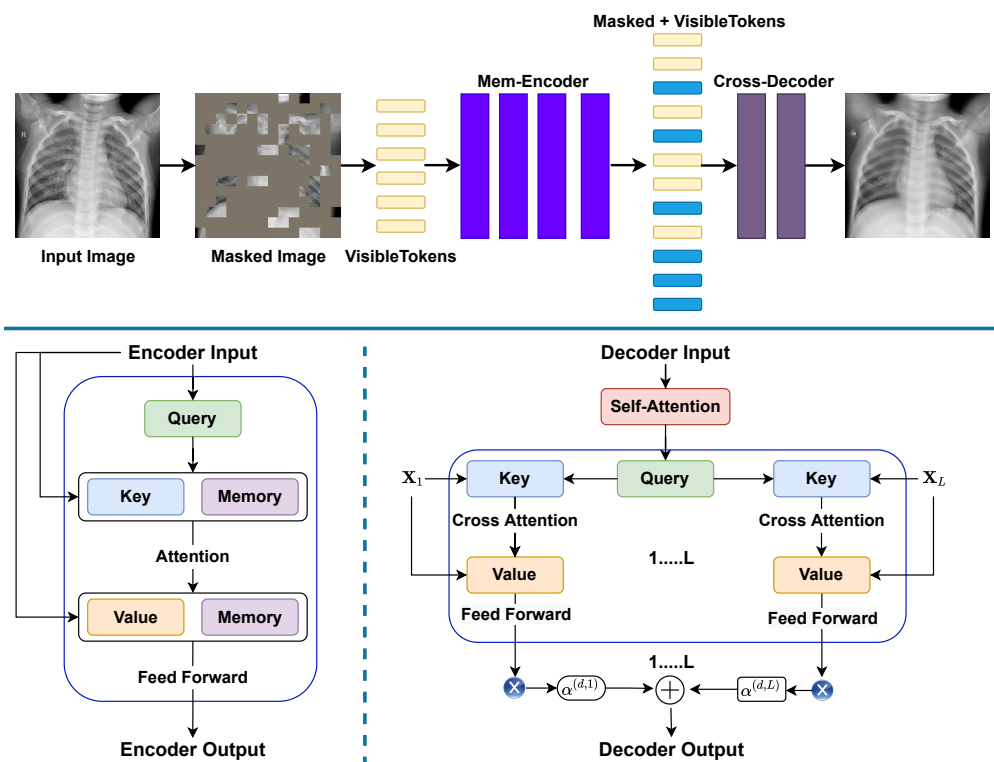


Figure 7.1: **Top:** overall MemMC-MAE framework. Yellow tokens indicate the unmasked visible patches, and blue tokens indicate the masked patches. Our memory-augmented transformer encoder only accepts the visible patches/tokens as input, and its output tokens are combined with dummy masked patches/tokens for the missing pixel reconstruction using our proposed multi-level cross-attentional transformer decoder. **Bottom-left:** proposed memory-augmented self-attention operator for the transformer encoder, and **bottom-right:** proposed multi-level cross-attention operator for the transformer decoder.

patterns of those patches, and all $|\mathcal{P}^{(v)}|$ encoded visible patches and $|\mathcal{P}| - |\mathcal{P}^{(v)}|$ dummy masked patches are used as the input of a new multi-level cross-attention decoder to reconstruct the image.

The training of MemMC-MAE is based on the minimisation of the mean squared error (MSE) loss between the input and reconstructed images at the pixels of the masked patches of the training images. The approach is evaluated on a testing set $\mathcal{T} = \{(\mathbf{x}, y, \mathbf{m})_i\}_{i=1}^{|\mathcal{T}|}$, where $y \in \mathcal{Y} = \{\text{normal}, \text{anomalous}\}$, and $\mathbf{m} \in \mathcal{M} \subset \{0, 1\}^{H \times W \times 1}$ denotes the segmentation mask of the lesion in the image \mathbf{x} . When testing, we also mask 75% of the image and the patch-wise reconstruction error indicates anomaly localisation, and the mean reconstruction error of all patches is used to detect image-wise anomaly. Below we provide details on the major contributions of MemMC-MAE, which

are the memory-augmented transformer encoder that stores the long-term normality patterns of the training samples, and the new multi-level cross-attentional transformer decoder to leverage the correlation of features from the encoder to reconstruct the missing normal pixels.

Memory-augmented Transformer Encoder (Fig. 7.1 - bottom left)

We modify the encoder from the transformer with our a novel memory-augmented self-attention, by extending the keys and values of the self-attention operation with learnable memory matrices that store normality patterns, which are updated via back-propagation. To this end, the proposed self-attention (SA) module for layer $l \in \{0, \dots, L - 1\}$ is defined as:

$$\mathbf{X}^{(l+1)} = f_{SA}(\mathbf{W}_Q^{(l)}\mathbf{X}^{(l)}, [\mathbf{W}_K^{(l)}\mathbf{X}^{(l)}, \mathbf{M}_K^{(l)}], [\mathbf{W}_V^{(l)}\mathbf{X}^{(l)}, \mathbf{M}_V^{(l)}]), \quad (7.1)$$

where $\mathbf{X}^{(0)}$ is the encoder input matrix containing $|\mathcal{P}^{(v)}|$ patch tokens formed from the visible image patches transformed through the linear projection $\mathbf{W}^{(0)}$, with $|\mathcal{P}^{(v)}|$ being the number of visible tokens/patches, $\mathbf{X}^{(l)}, \mathbf{X}^{(l+1)}$ are the input and output of layer l , $\mathbf{W}_Q^{(l)}, \mathbf{W}_K^{(l)}, \mathbf{W}_V^{(l)}$ are the linear projections of the encoder’s layer l for query, key and value of the self-attention operator, respectively, and $\mathbf{M}_K^{(l)}, \mathbf{M}_V^{(l)}$ are the layer l learnable memory matrices that are concatenated with $\mathbf{W}_K^{(l)}\mathbf{X}^{(l)}$ and $\mathbf{W}_V^{(l)}\mathbf{X}^{(l)}$ using the operator $[\cdot, \cdot]$. The self-attention operator $f_{SA}(\cdot)$ follows the standard ViT [53] and transformer [227], which computes a weighted sum of value vectors according to the cosine similarity distribution between query and key. Such memory-augmented self-attention aims to store normal patterns that are not encoded in the feature $\mathbf{X}^{(l)}$, forcing the decoder to reconstruct anomalous input patches into normal output patches during testing.

Multi-level Cross-Attention Transformer Decoder (Fig. 7.1 - bottom right).

Our transformer decoder computes the cross-attention operation using the outputs from all encoder layers and the decoder layer output from the self-attention operator (see Fig. 7.1 - Bottom right). More formally, the layer $d \in \{0, \dots, D - 1\}$ of our decoder outputs

$$\mathbf{Y}^{(d+1)} = \sum_{l=1}^L \alpha^{(d,l)} \times f_{SA}(f_{SA}(\mathbf{Y}^{(d)}, \mathbf{Y}^{(d)}, \mathbf{Y}^{(d)}), \mathbf{W}_K^{(d)}\mathbf{X}^{(l)}, \mathbf{W}_V^{(d)}\mathbf{X}^{(l)}), \quad (7.2)$$

where $\mathbf{Y}^{(d)}$ and $\mathbf{Y}^{(d+1)}$ represent the input and output of the decoder layer d containing $|\mathcal{P}|$ tokens (i.e., $|\mathcal{P}^{(v)}|$ tokens from the visible patches of the encoder and $|\mathcal{P}| - |\mathcal{P}^{(v)}|$ dummy tokens from the masked patches), $\mathbf{X}^{(l)}$ denotes the output from encoder layer

Methods	Publication	Covid-X (AUC)	Hyper-Kvasir (AUC)
DAE [60]	ICANN’11	0.557	0.705
OCGAN [174]	CVPR’18	0.612	0.813
F-anoGAN [199]	IPMI’17	0.669	0.907
ADGAN [140]	ISBI’19	0.659	0.913
MS-SSIM [33]	AAAI’22	0.634	0.917
PANDA [185]	CVPR’21	0.629	0.937
PaDiM [?]	ICPR’21	0.614	0.923
IGD [33]	AAAI’22	0.699	0.939
CCD+IGD* [221]	MICCAI’21	0.746	0.972
Ours		0.917	0.972

Table 7.1: **Anomaly detection AUC** test results on Covid-X and Hyper-Kvasir. CCD+IGD* [221] requires at least 2×longer training time than other approaches in the table because of a two-stage self-supervised pre-training and fine-tuning.

$l - 1$, and $\mathbf{W}_K^{(d)}$, $\mathbf{W}_V^{(d)}$ are the linear projections of the layer d of the decoder for the key and value of the self-attention operator, respectively. Note that all $|\mathcal{P}|$ input tokens for the decoder are attached with positional embeddings. The multi-level cross-attention results in (7.2) are fused together with a weighted sum operation using the weight $\alpha^{(l,d)}$, which is computed based on a linear projection layer and sigmoid function to control the weight of different layers’ cross-attention results, as in

$$\alpha^{(d,l)} = \sigma \left(\mathbf{W}_\alpha^{(d,l)} \left([f_{SA}(\mathbf{Y}^{(d)}, \mathbf{Y}^{(d)}, \mathbf{Y}^{(d)}), \mathbf{Y}^{(d+1)}] \right) \right), \quad (7.3)$$

where $\sigma(\cdot)$ is the sigmoid function, and $\mathbf{W}_\alpha^{(d,l)}$ denotes a learnable weight matrix. Such fusion mechanism enforces the correlation of multiple normal patterns in the image present at different levels of encoding information to contribute at different decoding layers by adjusting their relative importance using the self-attention output from $f_{SA}(\cdot)$ and cross-attention output $\mathbf{Y}^{(d+1)}$.

7.2.2 Anomaly Detection and Segmentation

We compute the anomaly score [33] with multi-scale structural similarity (MS-SSIM) [241]. The anomaly scores are pooled from 10 different random seeds for masking image patches with a fixed 75% masking ratio, which enables a more robust anomaly detection and localisation. The anomaly localisation mask is obtained by computing the mean MS-SSIM scores for all patches, and the anomaly detection relies on the mean MS-SSIM scores from the patches [33].

7.3 Experiments and Results

Datasets and Evaluation Measures

Two disease screening datasets are used in our experiments. We test anomaly detection on the CXR images of the Covid-X dataset [235], and both anomaly detection and localisation on the colonoscopy images of the Hyper-Kvasir dataset [21]. **Covid-X** [235] has a training set with 1,670 Covid-19 positive and 13,794 Covid-19 negative CXR images, but we only use the 13,794 Covid-19 negative CXR images for training. The test set contains 400 CXR images, consisting of 200 positive and 200 negative images, each image with size 299×299 pixels. **Hyper-Kvasir** is a large-scale public gastrointestinal dataset. The images were collected from the gastroscopy and colonoscopy procedures from Baerum Hospital in Norway, and were annotated by experienced medical practitioners. The dataset contains 110,079 images from unhealthy and healthy patients, out of which, 10,662 are labelled. Following [221], 2,100 normal images from ‘cecum’, ‘ileum’ and ‘bbps-2-3’ are selected, from which we use 1,600 for training and 500 for testing. The testing set also contains 1,000 anomalous images with their segmentation masks. Detection is assessed with area under the ROC curve (AUC) [33, 56, 60, 174], and localisation is evaluated with intersection over union (IoU) [33, 221, 229?].

Implementation Details

For the transformer, we follow ViT-B [53, 89] for designing the encoder and decoder, consisting of stacks of transformer blocks. Inspired by U-Net [269] for medical segmentation, we add residual connections to transfer information from earlier to later blocks for both the encoder and decoder. Each encoder block contains a memory-augmented self-attention block and an MLP block with LayerNorm (LN). Each decoder block contains a multi-level cross-attention block and an MLP block with LayerNorm (LN). We also adopt a linear projection layer after the encoder to match the different width between encoder and decoder [89]. We add positional embeddings (with the sine-cosine version) to both the encoder and decoder input tokens. RandomResizedCrop is used for data augmentation during training. Our method is trained for 2000 epochs in an end-to-end manner using the Adam optimiser [107] with a weight decay of 0.05 and a batch size of 256. The learning rate is set to $1.5e-3$. At the beginning, we warm up the training process for 5 epochs. The method is implemented in PyTorch [172] and run on an NVIDIA 3090 GPU. The overall training times is around 22 hours, and the mean inference time takes 0.21s per image.

Evaluation on Anomaly Detection on Covid-X and Hyper-Kvasir

We compare our method with nine competing UAD approaches: DAE [60], OCGAN [174], f-anogan [199], ADGAN [140], MS-SSIM autoencoder [33], PANDA [185],

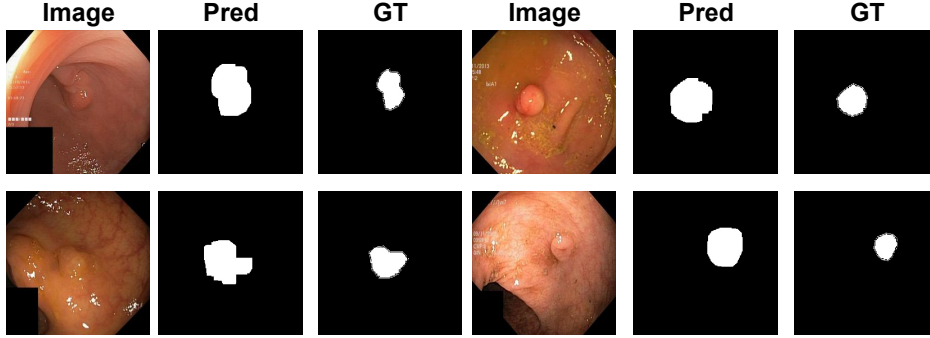


Figure 7.2: Segmentation results of our proposed method on Hyper-Kvasir [21], with our predictions (Pred) and ground truth annotations (GT).

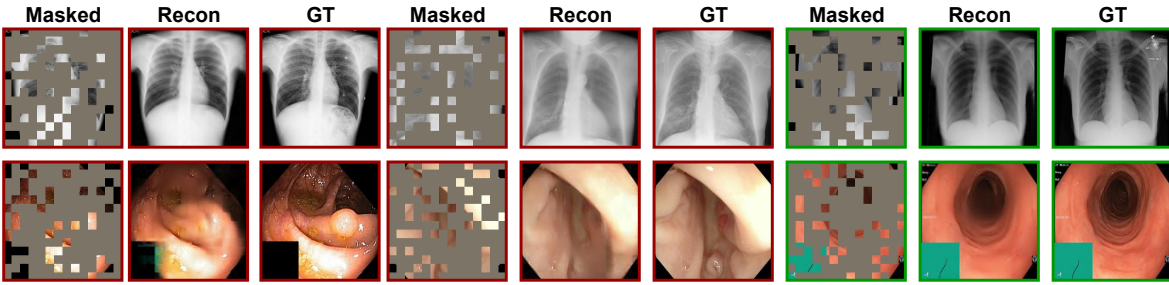


Figure 7.3: Reconstruction of testing images from Covid-X (Top) and Hyper-Kvasir (Bottom). For each triplet, we show the masked image (left), our MemMC-MAE reconstruction (middle), and the ground-truth (right). Normal testing images are marked with green boxes, and anomalous ones are marked with red boxes.

PaDiM [?], CCD [221] and IGD [33]. We apply the same experimental setup (i.e., image pre-processing, training strategy, evaluation methods) to these methods above as the one for our approach for fair comparison. The quantitative comparison results for anomaly detection are shown in Table 7.1 for both Covid-X and Hyper-Kvasir benchmarks. Our MemMC-MAE achieves the best AUC results on Covid-X and Hyper-Kvasir datasets with 91.7% and 97.2%, respectively. On Covid-X, our result outperforms all competing methods by a large margin with an improvement of 17.1% over the second best approach. For Hyper-Kvasir, our result is on par with the best result in the field produced by CCD+IGD [221], which has a training time $2\times$ longer than our approach.

Evaluation on Anomaly Localisation on Hyper-Kvasir

We compare our anomaly localisation results on Table 7.3 with four recently proposed UAD baselines: IGD [33], PaDiM [?], CCD [221] and CAVGA- R_u [229]. The results

MAE	Mem-Enc	MC-Dec	AUC - Covid
✓			0.799
✓	✓		0.862
✓	✓	✓	0.917

Table 7.2: **Ablation study** on Covid-X of the encoder’s memory-augmented operator (Mem-Enc) and the decoder’s multi-level cross-attention (MC-Dec).

Methods	Localisation - IoU
IGD [33]	0.276
PaDiM [?]]	0.341
CAVGA- R_u [229]	0.349
CCD + IGD [221]	0.372
Ours	0.419

Table 7.3: **Anomaly localisation:** Mean IoU test results on Hyper-Kvasir on 5 groups of 100 images.

of these methods on Table 7.3 are from [221]. Following [221], we randomly sample five groups of 100 anomalous images from the test set and compute the mean segmentation IoU. The proposed MemMC-MAE surpasses IGD, PaDiM, CAVGA- R_u and CCD by a minimum of 4.7% and a maximum of 14.3% IoU, illustrating the effectiveness of our model in localising anomalous tissues.

Visualisation of predicted segmentation.

The visualisation of polyp segmentation results of MemMC-MAE on Hyper-Kvasir [21] is shown in Fig. 7.2. Notice that our model can accurately segment colon polyps of various sizes and shapes.

Visualisation of Reconstructed Images

Figure 9.3 shows the reconstructions produced by MemMC-MAE on Covid-X (Top) and Hyper-Kvasir (Bottom) testing images. Notice that our method can effectively reconstruct the anomalous images with polyps/covid as normal images by automatically removing the polyps or blurring the anomalous regions, leading to larger reconstruction errors for those anomalies. The normal images are accurately reconstructed with smaller reconstruction errors than the anomalous images.

Ablation Study

Tab. 11.5 shows the contribution of each component of our proposed method on Covid-X testing set. The baseline MAE [89] achieves 79.9% AUC. Our method obtains a significant performance gain by adding the memory-augmented self-attention operator to the transformer encoder (Mem-Enc). Adding the proposed multi-level cross-attention operator into the decoder (MC-Dec) further boosts the performance by about 5% AUC.


7.4 Conclusion

We proposed a new UAD reconstruction method, called MemMC-MAE, for anomaly detection and localisation in medical images, which to the best of our knowledge, is the first UAD method based on MAE. MemMC-MAE introduced a novel memory-augmented self-attention operator for the MAE encoder and a new multi-level cross-attention for the MAE decoder to address the large reconstruction error of anomalous images that plague UAD reconstruction methods. The resulting anomaly detector showed SOTA anomaly detection and localisation accuracy on two public medical datasets. Despite the remarkable performance, the results can potentially improve if we use MemMC-MAE as a pre-training approach for other UAD methods [33, 221, 229?], which we plan to explore in the future.

Statement of Authorship

Title of Paper	Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning
Publication Status	<input checked="" type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	Published in International Conference on Computer Vision (ICCV) 2021.

Principal Author

Name of Principal Author (Candidate)	Yu Tian		
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper.		
Overall percentage (%)	90		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	09/03/2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Guansong Pang		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Yuanhong Chen		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Rajvinder Singh		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Johan W Verjans		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Please cut and paste additional co-author panels here as required.

Chapter 8

Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning

Abstract

Anomaly detection with weakly supervised video-level labels is typically formulated as a multiple instance learning (MIL) problem, in which we aim to identify snippets containing abnormal events, with each video represented as a bag of video snippets. Although current methods show effective detection performance, their recognition of the positive instances, i.e., rare abnormal snippets in the abnormal videos, is largely biased by the dominant negative instances, especially when the abnormal events are subtle anomalies that exhibit only small differences compared with normal events. This issue is exacerbated in many methods that ignore important video temporal dependencies. To address this issue, we introduce a novel and theoretically sound method, named Robust Temporal Feature Magnitude learning (RTFM), which trains a feature magnitude learning function to effectively recognise the positive instances, substantially improving the robustness of the MIL approach to the negative instances from abnormal videos. RTFM also adapts dilated convolutions and self-attention mechanisms to capture long- and short-range temporal dependencies to learn the feature magnitude more faithfully. Extensive experiments show that the RTFM-enabled MIL model (i) outperforms several state-of-the-art methods by a large margin on four benchmark data sets (ShanghaiTech, UCF-Crime, XD-Violence and UCSD-Peds) and (ii) achieves significantly improved subtle anomaly discriminability and sample efficiency.

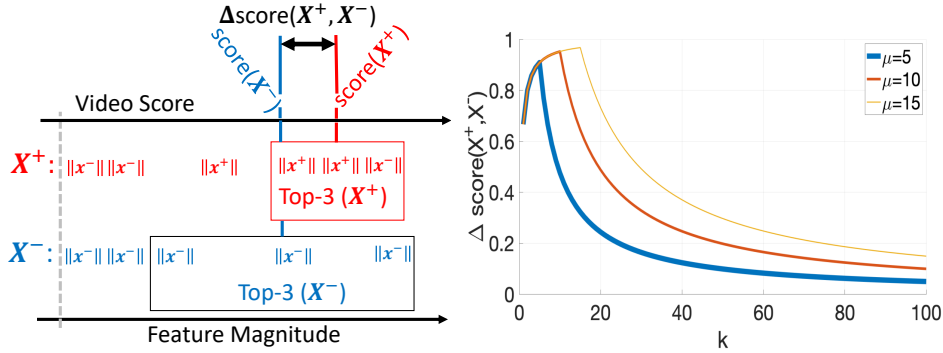


Figure 8.1: **RTFM** trains a feature magnitude learning function to improve the robustness of MIL approaches to normal snippets from abnormal videos, and detect abnormal snippets more effectively. **Left:** temporal feature magnitudes of abnormal and normal snippets ($\|\mathbf{x}^+\|$ and $\|\mathbf{x}^-\|$), from abnormal and normal videos (\mathbf{X}^+ and \mathbf{X}^-). Assuming that $\mu = 3$ denotes the number of abnormal snippets in the anomaly video, we can maximise the $\Delta\text{score}(\mathbf{X}^+, \mathbf{X}^-)$, which measures the difference between the scores of abnormal and normal videos, by selecting the top $k \leq \mu$ snippets with the largest temporal feature magnitude (the scores are computed with the mean of magnitudes of the top k snippets). **Right:** the $\Delta\text{score}(\mathbf{X}^+, \mathbf{X}^-)$ increases with $k \in [1, \mu]$ and then decreases for $k > \mu$, showing evidence that our proposed RTFM-enabled MIL model provides a better separation between abnormal and normal videos when $k \approx \mu$, even if there are a few normal snippets with large feature magnitudes.

8.1 Introduction

Video anomaly detection has been intensively studied because of its potential to be used in autonomous surveillance systems [87, 211, 245, 266]. The goal of video anomaly detection is to identify the time window when an anomalous event happened – in the context of surveillance, examples of anomaly are bullying, shoplifting, violence, etc. Although one-class classifiers (OCCs, also called unsupervised anomaly detection) trained exclusively with normal videos have been explored in this context [65, 87, 97, 145, 183, 184, 264], the best performing approaches explore a weakly-supervised setup using training samples with *video-level* label annotations of normal or abnormal [211, 245, 266]. This weakly-supervised setup targets a better anomaly classification accuracy at the expense of a relatively small human annotation effort, compared with OCC approaches.

One of the major challenges of weakly supervised anomaly detection is how to identify anomalous snippets from a whole video labelled as abnormal. This is due to two reasons, namely: 1) the majority of snippets from an abnormal video consist of

normal events, which can overwhelm the training process and challenge the fitting of the few abnormal snippets; and 2) abnormal snippets may not be sufficiently different from normal ones, making a clear separation between normal and abnormal snippets challenging. Anomaly detection trained with multiple-instance learning (MIL) approaches [211, 245, 262, 270] mitigates the issues above by balancing the training set with the same number of abnormal and normal snippets, where normal snippets are randomly selected from the normal videos and abnormal snippets are the ones with the top anomaly scores from abnormal videos. Although partly addressing the issues above, MIL introduces four problems: 1) the top anomaly score in an abnormal video may not be from an abnormal snippet; 2) normal snippets randomly selected from normal videos may be relatively easy to fit, which challenges training convergence; 3) if the video has more than one abnormal snippet, we miss the chance of having a more effective training process containing more abnormal snippets per video; and 4) the use of classification score provides a weak training signal that does not necessarily enable a good separation between normal and abnormal snippets. These issues are exacerbated even more in methods that ignore important temporal dependencies [65, 145, 245, 266].

To address the MIL problems above, we propose a novel method, named Robust Temporal Feature Magnitude (RTFM) learning. In RTFM, we rely on the temporal feature magnitude of video snippets, where features with low magnitude represent normal (i.e., negative) snippets and high magnitude features denote abnormal (i.e., positive) snippets. RTFM is theoretically motivated by the top- k instance MIL [124] that trains a classifier using k instances with top classification scores from the abnormal and normal videos, but in our formulation, we assume that the mean feature magnitude of abnormal snippets is larger than that of normal snippets, instead of assuming separability between the classification scores of abnormal and normal snippets [124]. RTFM solves the MIL issues above, as follows: 1) the probability of selecting abnormal snippets from abnormal videos increases; 2) the hard negative normal snippets selected from the normal videos will be harder to fit, improving training convergence; 3) it is possible to include more abnormal snippets per abnormal video; and 4) using feature magnitude to recognise positive instances is advantageous compared to MIL methods that use classification scores [124, 211], because it enables a stronger learning signal, particularly for the abnormal snippets that have a magnitude that can increase for the whole training process, and the feature magnitude learning can be jointly optimised with the MIL anomaly classification to enforce large margins between abnormal and normal snippets at both the feature representation space and the anomaly classification output space. Fig. 11.1 motivates RTFM, showing that the selection of the top- k features (based on their magnitude) can provide a better separation between abnormal and normal videos, when we have more than one abnormal snippet per abnormal video and the mean snippet feature magnitude of abnormal videos is larger than that of normal videos.

In practice, RTFM enforces large margins between the top k snippet features with largest magnitudes from abnormal and normal videos, which has theoretical guarantees to maximally separate abnormal and normal video representations. These top k snippet features from normal and abnormal videos are then selected to train a snippet classifier. To seamlessly incorporate long and short-range temporal dependencies within each video, we combine the learning of long and short-range temporal dependencies with a pyramid of dilated convolutions (PDC) [254] and a temporal self-attention module (TSA) [239]. We validate our RTFM on four anomaly detection benchmark data sets, namely ShanghaiTech [65], UCF-Crime [211], XD-Violence [245] and UCSD-Peds [123]. We show that our method outperforms the current SOTAs by a large margin on all benchmarks using different pre-trained features (i.e., C3D and I3D). We also show that our method achieves substantially better sample efficiency and subtle anomaly discriminability than popular MIL methods.

8.2 Related Work

Unsupervised Anomaly Detection. Traditional anomaly detection methods assume the availability of normal training data only and address the problem with one-class classification using handcrafted features [7, 152, 234, 263]. With the advent of deep learning, more recent approaches use the features from pre-trained deep neural networks [72, 101, 169, 205, 265]. Others apply constraints on the latent space of normal manifold to learn compact normality representations [2, 10, 12, 14, 34, 36, 45, 77, 141, 150, 153, 170, 175, 192, 196, 212, 221, 233, 268]. Alternatively, some approaches depend on data reconstruction using generative models to learn the representations of normal samples by (adversarially) minimising the reconstruction error [22, 56, 65, 100, 100, 154, 157, 159, 170, 187, 194, 196, 228, 249, 273]. These approaches assume that unseen anomalous videos/images often cannot be reconstructed well and consider samples of high reconstruction errors to be anomalies. However, due to the lack of prior knowledge of abnormality, these approaches can overfit the training data and fail to distinguish abnormal from normal events. Readers are referred to [165] for a comprehensive review of those anomaly detection approaches.

Weakly Supervised Anomaly Detection. Leveraging some labelled abnormal samples has shown substantially improved performance over the unsupervised approaches [137, 164, 167, 193, 211, 216, 245, 257, 258, 259]. However, large-scale frame-level label annotation is too expensive to obtain. Hence, current SOTA video anomaly detection approaches rely on weakly supervised training that uses cheaper video-level annotations. Sultani et al. [211] proposed the use of video-level labels and introduced the large-scale weakly-supervised video anomaly detection data set, UCF-Crime. Since then, this direction has attracted the attention of the research community [232, 245, 262].

Weakly-supervised video anomaly detection methods are mainly based on the MIL framework [211]. However, most MIL-based methods [211, 262, 270] fail to leverage abnormal video labels as they can be affected by the label noise in the positive bag caused by a normal snippet mistakenly selected as the top abnormal event in an anomaly video. To deal with this problem, Zhong et al. [266] reformulated this problem as a binary classification under noisy label problem and used a graph convolution neural (GCN) network to clear the label noise. Although this paper shows more accurate results than [211], the training of GCN and MIL is computationally costly, and it can lead to unconstrained latent space (i.e., normal and abnormal features can lie at any place of the feature space) that can cause unstable performance. By contrast, our method has trivial computational overheads compared to the original MIL formulation. Moreover, our method unifies the representation learning and anomaly score learning by an ℓ_2 -norm-based temporal feature ranking loss, enabling better separation between normal and abnormal feature representations, improving the exploration of weak labels compared to previous MIL methods [211, 231, 245, 262, 266, 270].

Temporal Dependency. Temporal Dependency has been explored in [65, 112, 137, 145, 245, 250, 266]. In anomaly detection, traditional methods [112, 250] convert consecutive frames into handcrafted motion trajectories to capture the local consistency between neighbouring frames. Diverse temporal dependency modelling methods have been used in deep anomaly detection approaches, such as stacked RNN [145], temporal consistency in future frame prediction [65], and convolution LSTM [137]. However, these methods capture short-range fixed-order temporal correlations only with single temporal scale, ignoring the long-range dependency from all possible temporal locations and the events with varying temporal length. GCN-based methods are explored in [245, 266] to capture the long-range dependency from snippets features, but they are inefficient and hard to train. By contrast, our proposed module combines PDC [254] and TSA [239] on the temporal dimension to seamlessly and efficiently incorporate both the long and short-range temporal dependencies into our temporal feature ranking loss.

8.3 The Proposed Method: RTFM

Our proposed robust temporal feature magnitude (RTFM) approach aims to differentiate between abnormal and normal snippets using weakly labelled videos for training. Given a set of weakly-labelled training videos $\mathcal{D} = \{(\mathbf{F}_i, y_i)\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{F} \in \mathcal{F} \subset \mathbb{R}^{T \times D}$ are pre-computed features (e.g., I3D [24] or C3D [222]) of dimension D from the T video snippets, and $y \in \mathcal{Y} = \{0, 1\}$ denotes the video-level annotation ($y_i = 0$ if \mathbf{F}_i is a normal video and $y_i = 1$ otherwise). The model used by RTFM is denoted by $r_{\theta, \phi}(\mathbf{F}) = f_{\phi}(s_{\theta}(\mathbf{F}))$ and returns a T -dimensional feature $[0, 1]^T$ representing the classification of the T video snippets into abnormal or normal, with the parameters θ, ϕ defined below. The training of this model comprises a joint optimisation of an end-

to-end multi-scale temporal feature learning, and feature magnitude learning and an RTFM-enabled MIL classifier training, with the loss

$$\min_{\theta, \phi} \sum_{i, j=1}^{|\mathcal{D}|} \ell_s(s_\theta(\mathbf{F}_i), (s_\theta(\mathbf{F}_j)), y_i, y_j) + \ell_f(f_\phi(s_\theta(\mathbf{F}_i)), y_i), \quad (8.1)$$

where $s_\theta : \mathcal{F} \rightarrow \mathcal{X}$ is the temporal feature extractor (with $\mathcal{X} \subset \mathbb{R}^{T \times D}$), $f_\phi : \mathcal{X} \rightarrow [0, 1]^T$ is the snippet classifier, $\ell_s(\cdot)$ denotes a loss function that maximises the separability between the top- k snippet features from normal and abnormal videos, and $\ell_f(\cdot)$ is a loss function to train the snippet classifier $f_\phi(\cdot)$ also using the top- k snippet features from normal and abnormal videos. Next, we discuss the theoretical motivation for our proposed RTFM, followed by a detailed description of the approach.

8.3.1 Theoretical Motivation of RTFM

Top- k MIL in [124] extends MIL to an environment where positive bags contain a minimum number of positive samples and negative bags also contain positive samples, but to a lesser extent, and it assumes that a classifier can separate positive and negative samples. Our problem is different because negative bags do not contain positive samples, and we do not make the classification separability assumption. Following the nomenclature introduced above, a temporal feature extracted from a video is denoted by $\mathbf{X} = s_\theta(\mathbf{F})$ in (9.1), where snippet features are represented by the rows \mathbf{x}_t of \mathbf{X} . An abnormal snippet is denoted by $\mathbf{x}^+ \sim P_x^+(\mathbf{x})$, and a normal snippet, $\mathbf{x}^- \sim P_x^-(\mathbf{x})$. An abnormal video \mathbf{X}^+ contains μ snippets drawn from $P_x^+(\mathbf{x})$ and $(T - \mu)$ drawn from $P_x^-(\mathbf{x})$, and a normal video \mathbf{X}^- has all T snippets sampled from $P_x^-(\mathbf{x})$.

To learn a function that can classify videos and snippets as normal or abnormal, we define a function that classifies a snippet using its magnitude (i.e., we use ℓ_2 norm to compute the feature magnitude), where instead of assuming classification separability between normal and abnormal snippets (as assumed in [124]), we make a milder assumption that $\mathbb{E}[\|\mathbf{x}^+\|_2] \geq \mathbb{E}[\|\mathbf{x}^-\|_2]$. This means that by learning the snippet feature from $s_\theta(\mathbf{F})$, such that normal ones have smaller feature magnitude than abnormal ones, we can satisfy this assumption. To enable such learning, we rely on an optimisation based on the mean feature magnitude of the top k snippets from a video [124], defined by

$$g_{\theta, k}(\mathbf{X}) = \max_{\Omega_k(\mathbf{X}) \subseteq \{\mathbf{x}_t\}_{t=1}^T} \frac{1}{k} \sum_{\mathbf{x}_t \in \Omega_k(\mathbf{X})} \|\mathbf{x}_t\|_2, \quad (8.2)$$

where $g_{\theta, k}(\cdot)$ is parameterised by θ to indicate its dependency on $s_\theta(\cdot)$ to produce \mathbf{x}_t , $\Omega_k(\mathbf{X})$ contains a subset of k snippets from $\{\mathbf{x}_t\}_{t=1}^T$ and $|\Omega_k(\mathbf{X})| = k$. The separability between abnormal and normal videos is denoted by

$$d_{\theta, k}(\mathbf{X}^+, \mathbf{X}^-) = g_{\theta, k}(\mathbf{X}^+) - g_{\theta, k}(\mathbf{X}^-). \quad (8.3)$$

For the theorem below, we define the probability that a snippet from $\Omega_k(\mathbf{X}^+)$ is abnormal with $p_k^+(\mathbf{X}^+) = \frac{\min(\mu, k)}{k+\epsilon}$, with $\epsilon > 0$ and from normal $\Omega_k(\mathbf{X}^-)$, $p_k^+(\mathbf{X}^-) = 0$. This definition means that it is likely to find an abnormal snippet within the top k snippets in $\Omega_k(\mathbf{X}^+)$, as long as $k \leq \mu$.

Theorem 8.3.1 (Expected Separability Between Abnormal and Normal Videos). *Assuming that $\mathbb{E}[\|\mathbf{x}^+\|_2] \geq \mathbb{E}[\|\mathbf{x}^-\|_2]$, where \mathbf{X}^+ has μ abnormal samples and $(T - \mu)$ normal samples, where $\mu \in [1, T]$, and \mathbf{X}^- has T normal samples. Let $D_{\theta, k}(\cdot)$ be the random variable from which the separability scores $d_{\theta, k}(\cdot)$ of (8.3) are drawn [124].*

1. If $0 < k < \mu$, then

$$0 \leq \mathbb{E}[D_{\theta, k}(\mathbf{X}^+, \mathbf{X}^-)] \leq \mathbb{E}[D_{\theta, k+1}(\mathbf{X}^+, \mathbf{X}^-)].$$

2. For a finite μ , then

$$\lim_{k \rightarrow \infty} \mathbb{E}[D_{\theta, k}(\mathbf{X}^+, \mathbf{X}^-)] = 0.$$

Proof.

$$\begin{aligned} \mathbb{E}[D_{\theta, k}(\mathbf{X}^+, \mathbf{X}^-)] &= \mathbb{E}[g_{\theta, k}(\mathbf{X}^+)] - \mathbb{E}[g_{\theta, k}(\mathbf{X}^-)] \\ &= p_k^+(\mathbf{X}^+) \mathbb{E}[\|\mathbf{x}^+\|_2] + p_k^-(\mathbf{X}^+) \mathbb{E}[\|\mathbf{x}^-\|_2] - \mathbb{E}[\|\mathbf{x}^-\|_2] \end{aligned} \quad (8.4)$$

1. Trivial given that $\mathbb{E}[\|\mathbf{x}^+\|_2] \geq \mathbb{E}[\|\mathbf{x}^-\|_2]$ and that $p_{k+1}^+(\mathbf{X}^+) > p_k^+(\mathbf{X}^+)$ for $0 < k < \mu$
2. Trivial given that as μ is finite, $\lim_{k \rightarrow \infty} p_k^+(\mathbf{X}^+) = 0$.

□

Therefore, the first part of this theorem means that as we include more samples in the top k snippets of the abnormal video, the separability between abnormal and normal video tends to increase (even if it includes a few normal samples) as long as $k \leq \mu$. The second part of the theorem means that as we include more than μ top instances, the abnormal and normal video scores become indistinguishable because of the overwhelming number of negative samples both in the positive and negative bags. Both points are shown in Fig. 11.1, where $\text{score}(\mathbf{X}) = g_{\theta, k}(\mathbf{X})$, $\Delta \text{score}(\mathbf{X}^+, \mathbf{X}^-) = d_{\theta, k}(\mathbf{X}^+, \mathbf{X}^-)$, and $\epsilon = 0.4$ to compute $p_k^+(\mathbf{X}^+)$. This theorem suggests that by maximising the separability of the top- k temporal feature snippets from abnormal and normal videos (for $k \leq \mu$), we can facilitate the classification of anomaly videos and snippets. It also suggests that the use of the top- k features to train the snippet classifier allows for a more effective training given that the majority of the top- k samples in the abnormal video will be abnormal and that we will have a balanced training using the top- k hardest normal snippets. The final consideration is that because we use just the top- k samples per video, our method is efficiently optimised with a relatively small amount of training samples.

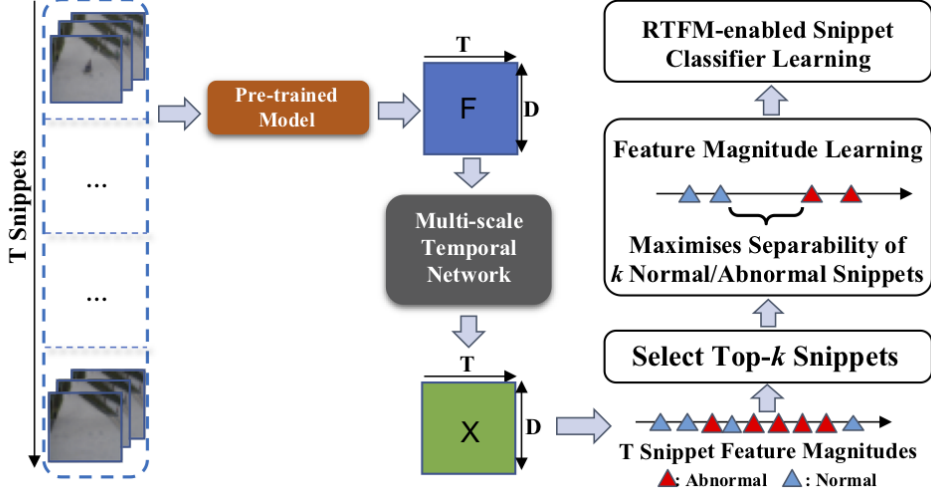


Figure 8.2: Our proposed RTFM receives a $T \times D$ feature matrix \mathbf{F} extracted from a video containing T snippets. Then, MTN captures the long and short-range temporal dependencies between snippet features to produce $\mathbf{X} = s_{\theta}(\mathbf{F})$. Next, we maximise the separability between abnormal and normal video features and train a snippet classifier using the top- k largest magnitude feature snippets from abnormal and normal videos.

8.3.2 Multi-scale Temporal Feature Learning

Inspired by the attention techniques used in video understanding [132, 239], our proposed multi-scale temporal network (MTN) captures the multi-resolution local temporal dependencies and the global temporal dependencies between video snippets (as shown in Fig. 8.3). MTN uses a pyramid of dilated convolutions over the time domain to learn multi-scale representations for video snippets. Dilated convolution is usually applied in the spatial domain with the goal of expanding the receptive field without losing resolution [254]. Here we propose to use dilated convolutions over the temporal dimension as it is important to capture the multi-scale temporal dependencies of neighbouring video snippets for anomaly detection.

MTN learns the multi-scale temporal features from the pre-computed fetures $\mathbf{F} = [\mathbf{f}_d]_{d=1}^D$. Then given the feature $\mathbf{f}_d \in \mathbb{R}^T$, the 1-D dilated convolution operation with kernel $\mathbf{W}_{k,d}^{(l)} \in \mathbb{R}^W$ with $k \in \{1, \dots, D/4\}$, $d \in \{1, \dots, D\}$, $l \in \{\text{PDC}_1, \text{PDC}_2, \text{PDC}_3\}$, and W denoting the filter size, is defined by

$$\mathbf{f}_k^{(l)} = \sum_{d=1}^D \mathbf{W}_{k,d}^{(l)} *^{(l)} \mathbf{f}_d, \quad (8.5)$$

where $*^{(l)}$ represents the dilated convolution operator indexed by l , $\mathbf{f}_k^{(l)} \in \mathbb{R}^T$ represents

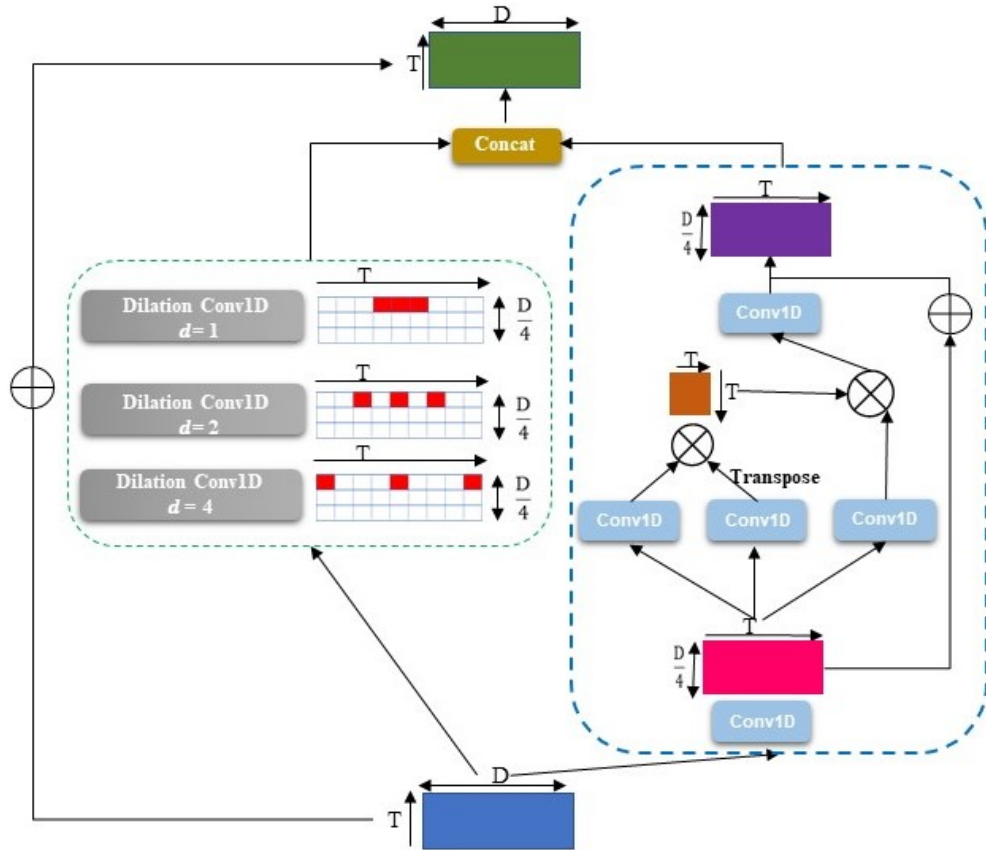


Figure 8.3: Our proposed MTN consists of two modules. The module on the left uses the pyramid dilated convolutions to capture the local consecutive snippets dependency over different temporal scales. The module on the right relies on a self-attention network to compute the global temporal correlations. The features from the two modules are concatenated to produce the MTN output.

the output features after applying the dilated convolution over the temporal dimension. The dilation factors for $\{PDC_1, PDC_2, PDC_3\}$ are $\{1, 2, 4\}$, respectively.

The global temporal dependencies between video snippets is achieved with a self-attention module, which has shown promising performance on capturing the long-range spatial dependency on video understanding [239], image classification [265] and object detection [178]. Motivated by the previous works using GCN to model global temporal information [245, 266], we re-formulate the spatial self-attention technique to work on the time dimension and capture global temporal context modelling. In detail, we aim to produce an attention map $\mathbf{M} \in \mathbb{R}^{T \times T}$ that estimates the pairwise correlation between snippets. Our temporal self-attention (TSA) module first uses a

1×1 convolution to reduce the spatial dimension from $\mathbf{F} \in \mathbb{R}^{T \times D}$ to $\mathbf{F}^{(c)} \in \mathbb{R}^{T \times D/4}$ with $\mathbf{F}^{(c)} = \text{Conv}_{1 \times 1}(\mathbf{F})$. We then apply three separate 1×1 convolution layers to $\mathbf{F}^{(c)}$ to produce $\mathbf{F}^{(c1)}, \mathbf{F}^{(c2)}, \mathbf{F}^{(c3)} \in \mathbb{R}^{T \times D/4}$, as in $\mathbf{F}^{(ci)} = \text{Conv}_{1 \times 1}(\mathbf{F}^{(c)})$ for $i \in \{1, 2, 3\}$. The attention map is then built with $\mathbf{M} = (\mathbf{F}^{(c1)}) (\mathbf{F}^{(c2)})^\top$, which produces $\mathbf{F}^{(c4)} = \text{Conv}_{1 \times 1}(\mathbf{M} \mathbf{F}^{(c3)})$.

A skip connection is added after this final 1×1 convolutional layer, as in

$$\mathbf{F}^{(\text{TSA})} = \mathbf{F}^{(c4)} + \mathbf{F}^{(c)}. \quad (8.6)$$

The output from the MTN is formed with a concatenation of the outputs from the PDC and MTN modules $\bar{\mathbf{F}} = [\mathbf{F}^{(l)}]_{l \in \mathcal{L}} \in \mathbb{R}^{T \times D}$, with $\mathcal{L} = \{\text{PDC}_1, \text{PDC}_2, \text{PDC}_3, \text{TSA}\}$. A skip connection using the original features \mathbf{F} produces the final temporal feature representation $\mathbf{X} = s_\theta(\mathbf{F}) = \bar{\mathbf{F}} + \mathbf{F}$, where the parameter θ comprises the weights for all convolutions described in this section.

8.3.3 Feature Magnitude Learning

Using the theory introduced in Sec. 8.3.1, we propose a loss function to model $s_\theta(\mathbf{F})$ in (9.1), where the top k largest snippet feature magnitudes from normal videos are minimised and the top k largest snippet feature magnitudes from abnormal videos are maximised. More specifically, we propose the following loss $\ell_s(\cdot)$ from (9.1) that maximises the separability between normal and abnormal videos:

$$\ell_s(s_\theta(\mathbf{F}_i), s_\theta(\mathbf{F}_j), y_i, y_j) = \begin{cases} \max\left(0, m - d_{\theta, k}(\mathbf{X}_i, \mathbf{X}_j)\right) & , \text{if } y_i = 1, y_j = 0 \\ 0 & , \text{otherwise} \end{cases} \quad (8.7)$$

where m is a pre-defined margin, $\mathbf{X}_i = s_\theta(\mathbf{F}_i)$ is the abnormal video feature (similarly for \mathbf{X}_j for a normal video), and $d_{\theta, k}(\cdot)$ represents separability function defined in (8.3) that computes the difference between the score of the top k instances, from $g_{\theta, k}(\cdot)$ in (8.2), of the abnormal and normal videos.

8.3.4 RTFM-enabled Snippet Classifier Learning

To learn the snippet classifier, we train a binary cross-entropy-based classification loss function using the set $\Omega_k(\mathbf{X})$ that contains the k snippets with the largest ℓ_2 -norm features from $s_\theta(\mathbf{F})$ in (9.1). In particular, the loss $\ell_f(\cdot)$ from (9.1) is defined as

$$\ell_f(f_\phi(s_\theta(\mathbf{F})), y) = \sum_{\mathbf{x} \in \Omega_k(\mathbf{X})} -(y \log(f_\phi(\mathbf{x})) + (1 - y) \log(1 - f_\phi(\mathbf{x}))), \quad (8.8)$$

where $\mathbf{x} = s_\theta(\mathbf{f})$. Note that following [211], $\ell_f(\cdot)$ is accompanied by the temporal smoothness and sparsity regularisation, with the temporal smoothness defined as $(f_\phi(s_\theta(\mathbf{f}_t)) - f_\phi(s_\theta(\mathbf{f}_{t-1})))^2$ to enforce similar anomaly score for neighbouring snippets, while the sparsity regularisation defined as $\sum_{t=1}^T |f_\phi(s_\theta(\mathbf{f}_t))|$ to impose a prior that abnormal events are rare in each abnormal video.

8.4 Experiments

8.4.1 Data Sets and Evaluation Measure

Our model is evaluated on four multi-scene benchmark datasets, created for the weakly supervised video anomaly detection task: ShanghaiTech [65], UCF-Crime [211], XD-Violence [245] and UCSD-Peds [250].

UCF-Crime is a large-scale anomaly detection data set [211] that contains 1900 untrimmed videos with a total duration of 128 hours from real-world street and indoor surveillance cameras. Unlike the static backgrounds in ShanghaiTech, UCF-Crime consists of complicated and diverse backgrounds. Both training and testing sets contain the same number of normal and abnormal videos. The data set covers 13 classes of anomalies in 1,610 training videos with video-level labels and 290 test videos with frame-level labels.

XD-Violence is a recently proposed large-scale multi-scene anomaly detection data set, collected from real life movies, online videos, sport streaming, surveillance cameras and CCTVs [245]. The total duration of this data set is over 217 hours, containing 4754 untrimmed videos with video-level labels in the training set and frame-level labels in the testing set. It is currently the largest publicly available video anomaly detection data set.

ShanghaiTech is a medium-scale data set from fixed-angle street video surveillance. It has 13 different background scenes and 437 videos, including 307 normal videos and 130 anomaly videos. The original data set [65] is a popular benchmark for the anomaly detection task that assumes the availability of normal training data. Zhong et al. [266] reorganised the data set by selecting a subset of anomalous testing videos into training data to build a weakly supervised training set, so that both training and testing sets cover all 13 background scenes. We use exactly the same procedure as in [266] to convert ShanghaiTech for the weakly supervised setting.

UCSD-Peds is a small-scale dataset combined by two sub-datasets – Ped1 with 70 videos and Peds2 with 28 videos. Previous work [88, 266] re-formulate the dataset for weakly supervised anomaly detection by randomly selecting 6 anomaly videos and 4 normal videos into the train set, with the remaining as test set. We report the mean results over 10 times of this process.

Evaluation Measure. Similarly to previous papers [56, 65, 211, 232, 262], we use

the frame-level area under the ROC curve (AUC) as the evaluation measure for all data sets. Moreover, following [245], we also use average precision (AP) as the evaluation measure for the XD-Violence data set. Larger AUC and AP values indicate better performance. Some recent studies [75, 182] recommend using the region-based detection criterion (RBDC) and the track-based detection criterion (TBDC) to complement the AUC measure, but these two measures are inapplicable in the weakly-supervised setting. Thus, we focus on the AUC and AP measures.

8.4.2 Implementation Details

Following [211], each video is divided into 32 video snippets, i.e., $T = 32$. For all experiments, we set the margin $m = 100$, $k = 3$ in (9.4). The three FC layers described in the model (Sec. 8.3) have 512, 128 and 1 nodes, where each of those FC layers is followed by a ReLU activation function and a dropout function with a dropout rate of 0.7. The 2048D and 4096D features are extracted from the 'mix_5c' and 'fc_6' layer of the pre-trained I3D [105] or C3D [104] network, respectively. In MTN, we set the pyramid dilate rate as 1, 2 and 4, and we use the 3×1 Conv1D for each dilated convolution branch. For the self-attention block, we use a 1×1 Conv1D.

Our RTFM method is trained in an end-to-end manner using the Adam optimiser [107] with a weight decay of 0.0005 and a batch size of 64 for 50 epochs. The learning rate is set to 0.001 for ShanghaiTech and UCF-Crime, and 0.0001 for XD-Violence. Each mini-batch consists of samples from 32 randomly selected normal and abnormal videos. The method is implemented using PyTorch [172]. For all baselines, we use the published results with the same backbone as ours. For a fair comparison, we use the same benchmark setup as in [211, 245, 266].

8.4.3 Results on ShanghaiTech

The frame-level AUC results on ShanghaiTech are shown in Tab. 8.1. Our method RTFM achieves superior performance when compared with previous SOTA unsupervised learning methods [65, 87, 145, 170, 255] and weakly-supervised approaches [231, 262, 266]. With I3D-RGB features, our model obtains the best AUC result on this data set: 97.21%. Using the same I3D-RGB features, our RTFM-enabled MIL method outperforms current SOTA MIL-based methods [211, 231, 262] by 10% to 14%. Our model outperforms [231] by more than 5% even though they rely on a more advanced feature extractor (i.e., I3D-RGB and I3D Flow). These results demonstrate the gains achieved from our proposed feature magnitude learning.

Our method also outperforms the GCN-based weakly-supervised method [266] by 11.7%, which indicates that our MTN module is more effective at capturing temporal dependencies than GCN. Additionally, considering the C3D-RGB features, our model

achieves the SOTA AUC of 91.51%, significantly surpassing the previous methods with C3D-RGB by a large margin.

Supervision	Method	Feature	AUC(%)
Unsupervised	Conv-AE [87]	-	60.85
	Stacked-RNN [145]	-	68.00
	Frame-Pred [65]	-	73.40
	Mem-AE [56]	-	71.20
	MNAD [170]	-	70.50
	VEC [255]	-	74.80
Weakly Supervised	GCN-Anomaly [266]	C3D-RGB	76.44
	GCN-Anomaly [266]	TSN-Flow	84.13
	GCN-Anomaly [266]	TSN-RGB	84.44
	Zhang et al. [262]	I3D-RGB	82.50
	Sultani et al.* [211]	I3D RGB	85.33
	AR-Net [231]	I3D Flow	82.32
	AR-Net [231]	I3D-RGB	85.38
	AR-Net [231]	I3D-RGB & I3D Flow	91.24
	Ours	C3D-RGB	91.51
	Ours	I3D-RGB	97.21

Table 8.1: Comparison of frame-level AUC performance with other SOTA un/weakly-supervised methods on ShanghaiTech. * indicates we retrain the method in [211] using I3D features. Best result in **red** and second best in **blue**.

8.4.4 Results on UCF-Crime

The AUC results on UCF-Crime are shown in Tab. 8.2. Our method outperforms all previous unsupervised learning approaches [87, 145, 209, 233]. Remarkably, using the same I3D-RGB features, our method also outperforms current SOTA MIL-based methods, Sultani et al. [211] by 8.62%, Zhang et al. [262] by 5.37%, Zhu et al. [270] by 5.03% and Wu et al. [245] by 1.59%. Zhong et al. [266] use a computationally costly alternating training scheme to achieve an AUC of 82.12%, while our method utilises an efficient end-to-end training scheme and outperforms their approach by 1.91%. Our method also surpasses the current SOTA unsupervised methods, BODS and GODS [233], by at least 13%. Considering the C3D features, our method surpasses the previous weakly supervised methods by a minimum 2.95% and a maximum 7.87%, indicating the effectiveness of our RTFM approach regardless of the backbone structure.

8.4.5 Results on XD-Violence

XD-Violence is a recently released data set, on which few results have been reported, as displayed in Tab. 8.3. Our approach surpasses all unsupervised learning approaches by a minimum of 27.03% in AP. Comparing with SOTA weakly-supervised methods [211,

Supervision	Method	Feature	AUC (%)
Unsupervised	SVM Baseline	-	50.00
	Conv-AE [87]	-	50.60
	Sohrab et al. [209]	-	58.50
	Lu et al. [143]	C3D RGB	65.51
	BODS [233]	I3D RGB	68.26
	GODS [233]	I3D RGB	70.46
Weakly Supervised	Sultani et al. [211]	C3D RGB	75.41
	Sultani et al.* [211]	I3D RGB	77.92
	Zhang et al. [262]	C3D RGB	78.66
	Motion-Aware [270]	PWC Flow	79.00
	GCN-Anomaly [266]	C3D RGB	81.08
	GCN-Anomaly [266]	TSN Flow	78.08
	GCN-Anomaly [266]	TSN RGB	82.12
	Wu et al. [245]	I3D RGB	82.44
	Ours	C3D RGB	83.28
Ours	I3D RGB	84.30	

Table 8.2: Frame-level AUC performance on UCF-Crime. * indicates we retrain the method in [211] using I3D features. Best result in **red** and second best in **blue**.

[245], our method is 2.4% and 2.13% better than Wu et al. [245] and Sultani et al. [211], using the same I3D features. With the C3D features, our RTFM achieves the best 75.89% AUC when compared with the MIL baseline by Sultani et al. [211]. The consistent superiority of our method reinforces the effectiveness of our proposed feature magnitude learning method in enabling the MIL-based anomaly classification.

Supervision	Method	Feature	AP(%)
Unsupervised	SVM baseline	-	50.78
	OCSVM [203]	-	27.25
	Hasan et al. [87]	-	30.77
Weakly Supervised	Sultani et al. [211]	C3D RGB	73.20
	Sultani et al.* [211]	I3D RGB	75.68
	Wu et al. [245]	I3D RGB	75.41
	Ours	C3D RGB	75.89
	Ours	I3D RGB	77.81

Table 8.3: Comparison of AP performance with other SOTA un/weakly-supervised methods on XD-Violence. * indicates we retrain the method in [211] using I3D features. Best result in **red** and second best in **blue**.

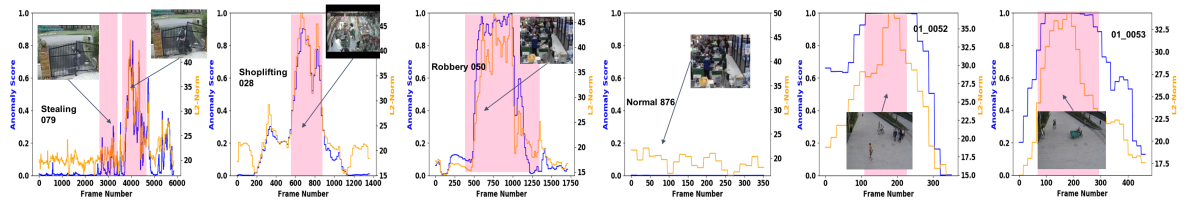


Figure 8.4: Anomaly scores and feature magnitude values of our method on UCF-Crime (*stealing079,shoplifting028, robbery050 normal876*), and ShanghaiTech (*01_0052, 01_0053*) test videos. Pink areas indicate the manually labelled abnormal events.

8.4.6 Results on UCSD-Peds

We showed the result on UCSD-Ped2 in Tab. 8.4, with TSN-Gray and I3D-RGB features, respectively. Our approach surpasses the previous SOTA [266] by a large 3.2% with the same TSN-Gray features. Finally, we achieves the best 98.6% mean AUC, surpassing Sultani et al. [211] by 6.3%, using the same I3D features.

Method	Feature	AUC (%)
GCN-Anomaly [266]	TSN-Flow	92.8
GCN-Anomaly [266]	TSN-Gray	93.2
Sultani et al.*[211]	I3D RGB	92.3
Ours	TSN-Gray	96.5
Ours	I3D-RGB	98.6

Table 8.4: Comparison of AUC performance with other SOTA weakly-supervised methods on UCSD Ped2. * indicates we retrain the method in [211] using I3D features. Best result in **red** and second best in **blue**.

8.4.7 Sample Efficiency Analysis

We investigate the sample efficiency of our method by looking into its performance w.r.t. the number of abnormal videos used for training on ShanghaiTech. We reduce the number of abnormal training videos from the original 63 videos down to 25 videos, with the normal training videos and test data fixed. The MIL method in [211] is used as a baseline. For a fair comparison, the same I3D features are used in both methods, and average AUC results ((computed from three runs using different random seeds)) are shown in Fig. 8.5. As expected, the performance of both our method and Sultani et al. [211] decreases with decreasing number of abnormal training videos, but the decreasing rate of our model is smaller that of than Sultani et al. [211], indicating the

robustness of our RTFM. Remarkably, our method using only 25 abnormal training videos outperforms [211] using all 63 abnormal videos by about 3%, i.e., although our method uses 60% less labelled abnormal training videos, it can still outperform Sultani et al. [211]. This is because RTFM performs better recognition of the positive instances in the abnormal videos, and as a result, it can leverage the same training data more effectively than a MIL-based approach [211]. Note that we retrain Sultani et al.’s method using the same I3D features.

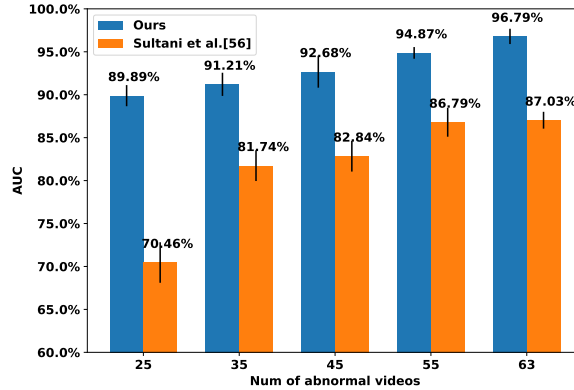


Figure 8.5: AUC w.r.t. the number of abnormal training videos.

8.4.8 Subtle Anomaly Discriminability

We also examine the ability of our method to detect subtle abnormal events on the UCF-Crime dataset, by studying the AUC performance on each individual anomaly class. The models are trained on the full training data and we use [211] as baseline, and results are shown in Fig. 8.6. Our model shows remarkable performance on human-centric abnormal events, even when the abnormality is very subtle. Particularly, our RTFM method outperforms Sultani et al. [211] in 8 human-centric anomaly classes (i.e., arson, assault, burglary, robbery, shooting, shoplifting, stealing, vandalism), significantly lifting the AUC performance by 10% to 15% in subtle anomaly classes such as burglary, shoplifting, vandalism. This superiority is supported the theoretical results of RTFM that guarantee a good separability of the positive and negative instances. For the arrest, fighting, road accidents and explosion classes, our method shows competitive performance to [211]. Our model is less effective in the abuse class because this class contains overwhelming human-centric abuse events in the training data but its testing videos contain animal abuse events only.

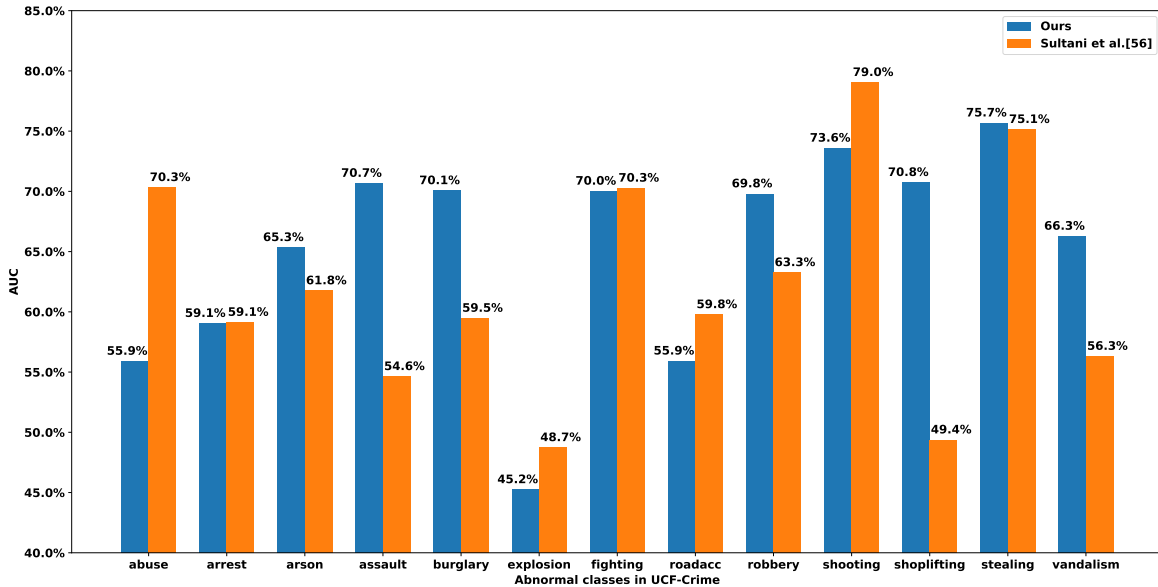


Figure 8.6: AUC results w.r.t. individual classes on UCF-Crime.

8.5 Computational Efficiency

We investigate if our system can run in real time. During inference, our method processes a 16-frame clip in 0.76 seconds on a Nvidia 2080Ti—this time includes the I3D extraction time. This indicates that our system can achieve good real-time detection in real-world applications.

8.5.1 Ablation Studies

We perform the ablation study on ShanghaiTech and UCF Crime with I3D features, as shown in Tab. 11.5, where the temporal feature mapping function s_θ is decomposed into PDC and TSA, and FM represents the feature magnitude learning from Sec. 8.3.3. The baseline model replaces PDC and TSA with a 1×1 convolutional layer and is trained with the original MIL approach as in [211]. The resulting baseline achieves only 85.96% AUC on ShanghaiTech and 77.32% AUC on UCF Crime (a result similar to the one in [211]). By adding PDC or TSA, the AUC performance is boosted to 89.21% and 91.73% on ShanghaiTech and 79.32% and 78.96% on UCF, respectively. When both PDC and TSA are added, the AUC result increases to 92.32% and 82.12% for the two datasets, respectively. This indicates that PDC and TSA contributes to the overall performance, and they also complement each other in capturing both long and short-range temporal relations. When adding only the FM module to the baseline,

Baseline	PDC	TSA	FM	AUC (%) - Shanghai	AUC (%) - UCF
✓				85.96	77.39
✓	✓			89.21	79.32
✓		✓		91.73	78.96
✓	✓	✓		92.32	82.12
✓			✓	92.99	81.28
✓		✓	✓	94.63	82.97
✓	✓		✓	93.91	82.58
✓	✓	✓	✓	97.21	84.30

Table 8.5: Ablation studies of our method on ShanghaiTech and UCF-Crime.

the AUC substantially increases by over 7% and 4% on ShanghaiTech and UCF Crime, respectively, indicating that our feature magnitude learning considerably improves over the original MIL method as it enables better exploitation of the labelled abnormal video data. Additionally, combining either PDC or TSA with FM helps further improve the performance. Then, the full model RTFM can achieve the best performance of 97.21% and 84.30% on the two datasets. An assumption made in theoretical motivation for RTFM is that the mean feature magnitudes for the top- k abnormal feature snippets is larger than the ones for normal snippets. We measure that on the testing videos of UCF-Crime and the mean magnitude of the top- k snippets from abnormal videos is 53.4 and for normal, it is 7.7. This shows empirically that our assumption for Theorem B.1.1 is valid and that RTFM can effectively maximise the separability between normal and abnormal video snippets. This is further evidenced by the mean classification scores of 0.85 for the abnormal snippets and 0.13 for the normal snippets.

8.5.2 Qualitative Analysis

In Fig. 8.4, we show the anomaly scores produced by our MIL anomaly classifier for diverse test videos from UCF-Crime and ShanghaiTech. Three anomalous videos and one normal video from UCF-Crime are used (*stealing079*, *shoplifting028*, *robbery050* and *normal876*). As illustrated by the ℓ_2 -norm value curve (i.e., orange curves), our FM module can effectively produce a small feature magnitude for normal snippets and a large magnitude for abnormal snippets. Furthermore, our model can successfully ensure large margins between the anomaly scores of the normal and abnormal snippets (i.e., blank and pink shadowed areas, respectively). Our model is also able to detect multiple anomalous events in one video (e.g., *stealing079*), which makes the problem more difficult. Also, for the anomalous events *stealing* and *shoplifting*, the abnormality is subtle and barely seen through the videos, but our model can still detect it. We also show the anomaly scores and feature magnitudes produced by our model for *01_0052* and *01_0053* from ShanghaiTech (last two figures in Fig. 8.4). Our model can effectively yield large anomaly scores for the anomalous event of vehicle entering in these two

scenes.

8.6 Conclusion

We introduced a novel method, named RTFM, that enables top- k MIL approaches for weakly supervised video anomaly detection. RTFM learns a temporal feature magnitude mapping function that 1) detects the rare abnormal snippets from abnormal videos containing many normal snippets, and 2) guarantees a large margin between normal and abnormal snippets. This improves the subsequent MIL-based anomaly classification in two major aspects: 1) our RTFM-enabled model learns more discriminative features that improve its ability in distinguishing complex anomalies (e.g., subtle anomalies) from hard negative examples; and 2) it also enables the MIL classifier to achieve significantly improved exploitation of the abnormal data. These two capabilities respectively result in better subtle anomaly discriminability and sample efficiency than current SOTA MIL methods. They are also the two main drivers for our model to achieve SOTA performance on all three large benchmarks.

Statement of Authorship

Title of Paper	Contrastive Transformer-based Multiple Instance Learning for Weakly Supervised Polyp Frame Detection
Publication Status	<input type="checkbox"/> Published <input checked="" type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	Accepted To International Conference on Medical Imaging Computing and Computer-Assisted Intervention (MICCAI) 2022

Principal Author

Name of Principal Author (Candidate)	Yu Tian
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper.
Overall percentage (%)	90
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	_____ Date 09/03/2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Guansong Pang
Contribution to the Paper	Discussion and writing the revision.
Signature	_____ Date 09/03/2022

Name of Co-Author	Fengbei Liu
Contribution to the Paper	Discussion and writing the revision.
Signature	_____ Date 09/03/2022

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Yuyuan Liu		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Chong Wang		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	11/03/2022

Name of Co-Author	Yuanhong Chen		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Johan W Verjans		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Please cut and paste additional co-author panels here as required.

Chapter 9

Contrastive Transformer-based Multiple Instance Learning for Weakly Supervised Polyp Frame Detection

Abstract

Current polyp detection methods from colonoscopy videos use exclusively normal (i.e., healthy) training images, which i) ignore the importance of temporal information in consecutive video frames, and ii) lack knowledge about the polyps. Consequently, they often have high detection errors, especially on challenging polyp cases (e.g., small, flat, or partially visible polyps). In this work, we formulate polyp detection as a weakly-supervised anomaly detection task that uses video-level labelled training data to detect frame-level polyps. In particular, we propose a novel convolutional transformer-based multiple instance learning method designed to identify abnormal frames (i.e., frames with polyps) from anomalous videos (i.e., videos containing at least one frame with polyp). In our method, local and global temporal dependencies are seamlessly captured while we simultaneously optimise video and snippet-level anomaly scores. A contrastive snippet mining method is also proposed to enable an effective modelling of the challenging polyp cases. The resulting method achieves a detection accuracy that is substantially better than current state-of-the-art approaches on a new large-scale colonoscopy video dataset introduced in this work.

9.1 Introduction and Background

Colonoscopy has become a vital exam for colorectal cancer (CRC) early diagnosis. This exam targets the early detection of polyps (a precursor of colon cancer), which

can improve survival rate by up to 95% [102, 181, 216]. During the procedure, doctors inspect the lower bowel with a scope to find polyps, but the quality of the exam depends on the ability of doctors to avoid mis-detections [181]. This can be alleviated by systems that automatically assist doctors detect frames containing polyps from colonoscopy videos. Nevertheless, accurate polyp detection is challenging due to the variable appearance, size and shape of colon polyps and their rare occurrence in an colonoscopy video.

One way to mitigate polyp detection challenges is with fully supervised training approaches, but given the expensive acquisition of fully labelled training sets, recent approaches have formulated the problem as an unsupervised anomaly detection (UAD) task [67, 139, 216, 221]. These UAD methods [139, 221] are trained with only normal training images and videos, and abnormal testing images and videos that contain polyps are detected as anomalous events. However, UAD approaches do not use training images or *snippets* (i.e., a set of consecutive video frames) containing polyps, so they are ineffective in recognising polyps of diverse characteristics, especially those that are small, partially visible, or irregularly shaped. As shown in a number of recent studies [167, 168, 211, 216, 217, 245], incorporating some knowledge about anomalies into the training of anomaly detectors has improved the detection accuracy of hard anomalies. For example, weakly-supervised video anomaly detection (WVAD) [211, 217, 245] relies on video-level labelled data to train detection models. The video-level labels only indicate whether the whole video contains anomalies or not, which is easier to acquire than fully-labelled datasets with frame-level annotations. The WVAD formulation is yet to be explored in the detection of polyps from colonoscopy, but it is of utmost importance because colonoscopy videos are often annotated with video-level labels in real-world datasets.

Most existing WVAD methods [73, 211, 217, 245, 266] rely on multiple instance learning (MIL), in which all snippets in a normal video are treated as normal snippets, while each abnormal video is assumed to have at least one abnormal snippet. This approach can utilise video-level labels to train an anomaly-informed detector to find anomalous frames, but MIL methods often fail to select rare abnormal snippets in anomalous videos, especially the challenging abnormal snippets that have subtle visual appearance differences from the normal ones (e.g., small and flat colon polyps or frames with partially visible polyps—see Fig. 9.2). Consequently, they perform poorly in detecting these subtle anomalous snippets. Moreover, the WVAD methods above are trained on individual images, ignoring the important temporal dependencies in colonoscopy videos that can be explored for a more stable polyp detection performance.

In this chapter, we introduce the first WVAD method specifically designed for detecting polyp frames from colonoscopy videos. Our method introduces a new contrastive snippet mining (CSM) algorithm to identify hard and easy normal and abnormal snippets. These snippets are further used to simultaneously optimise video and

snippet-level anomaly scores, which effectively reduces detection errors, such as misclassifying snippets with subtle polyps as normal ones, or normal snippets containing feces and water as abnormal ones. The exploration of global temporal dependency is also incorporated into our model with a transformer module, enabling a more stable anomaly classifier for colonoscopy videos. To resolve the poor modelling of local temporal dependency suffered by the transformer module [244], we also propose a convolutional transformer block to capture local correlations between neighbouring snippets. Our contributions are summarised as follows:

- To the best of our knowledge, this is the first work to tackle polyp detection from colonoscopy in a weakly supervised video anomaly detection manner.
- We propose a new transformer-based MIL framework that optimises anomaly scores in both snippet and video levels, resulting in more accurate anomaly scoring of polyp snippets.
- We introduce a new contrastive snippet mining (CSM) approach to identify hard and easy normal and abnormal snippets, where we pull the hard and easy snippets of the same class (i.e., normal or abnormal) together using a contrastive loss. This helps improve the robustness in detecting subtle polyp tissues and challenging normal snippets containing feces and water.
- We propose a new WVAD benchmark containing a large-scale diverse colonoscopy video dataset that combines several public colonoscopy datasets.

Our extensive empirical results show that our method achieves substantially better results than six state-of-the-art (SOTA) competing approaches on our newly proposed benchmark.

9.2 Method

Our method is trained with a set of weakly-labelled videos $\mathcal{D} = \{(\mathbf{F}_i, y_i)\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{F} \in \mathcal{F} \subset \mathbb{R}^{T \times D}$ represents pre-computed features (e.g., I3D [24]) of dimension D from T video snippets, and $y \in \mathcal{Y} = \{0, 1\}$ denotes the video-level annotation ($y_i = 0$ if \mathbf{F}_i is a normal video and $y_i = 1$ otherwise), with each video being equally divided into a fixed number of snippets. Our method aims to learn a convolutional transformer MIL anomaly classifier for the T snippets, as in $r_{\theta, \phi} : \mathcal{F} \rightarrow [0, 1]^T$, where this function is decomposed as $r_{\theta, \phi}(\mathbf{F}) = s_{\phi}(f_{\theta}(\mathbf{F}))$, with $f_{\theta} : \mathcal{F} \rightarrow \mathcal{X}$ being the transformer-based temporal feature encoder parameterised by θ (with $\mathcal{X} \subset \mathbb{R}^{T \times D}$) and $s_{\phi} : \mathcal{X} \rightarrow [0, 1]^T$ denoting the MIL anomaly classifier, parameterised by ϕ , to optimise snippet-level anomaly scores.

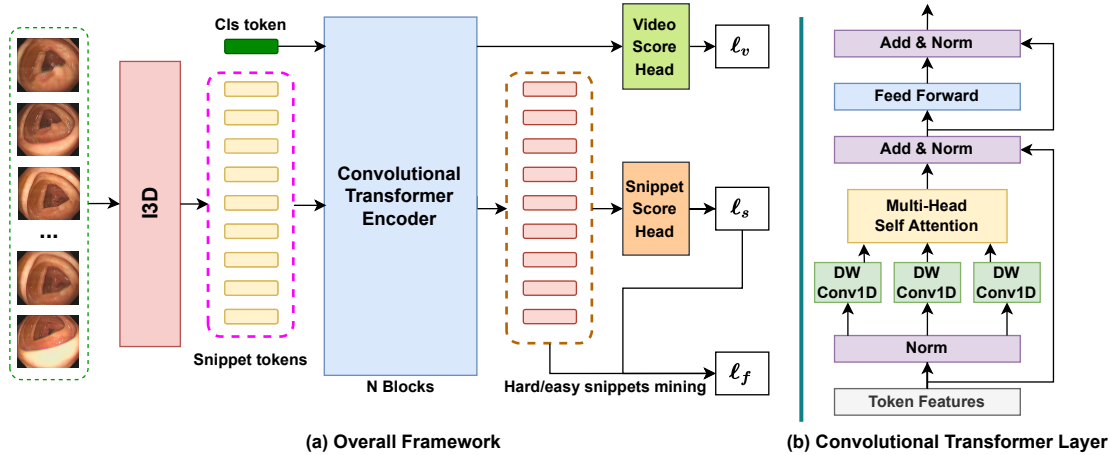


Figure 9.1: (a) The architecture of our method consists an I3D [24] snippet feature extractor and a Convolutional Transformer MIL Network. The I3D features are considered as snippet feature tokens to the transformer to predict snippet-wise anomaly scores using a snippet classifier. The CLs token is applied for a video classifier to predict if a video contains anomalies. The output features from the transformer are utilised to mine hard and easy snippets from normal and abnormal videos. The anomaly scores and hard/easy snippet representations are optimised by three proposed losses in (9.1). (b) The proposed Temporal Convolutional Transformer Layer replaces the linear projection with depthwise separable convolution (DW Conv1D) [38].

9.2.1 Convolutional Transformer MIL Network

Motivated by the recent success of transformer architectures in analysing the global context of images [53] and videos [6], we propose to use a transformer to model the temporal information between the snippets of colonoscopy videos. Standard transformer without convolution [53] cannot learn the local structure between adjacent snippets, which is important for modelling local temporal relations because adjacent snippets are often highly correlated [211, 221, 245]. Hence, we replace the linear token projection of the transformer by convolution operations. More specifically, we follow [244] and adopt the depth-wise separable 1D convolution [38] on the temporal dimension, as shown in Fig. 9.1(b). As shown in Fig. 9.1(a), the encoder comprises N convolutional transformer blocks that produce the final temporal feature representation $\mathbf{X} = f_{\theta}(\mathbf{F})$.

9.2.2 Transformer-based MIL Training

The training of our model comprises a joint optimisation of a transformer-based temporal feature learning, a contrastive snippet mining (CSM) that is used to train a

CSM-enabled MIL classifier, and a video-level classifier, with

$$\theta^*, \phi^*, \gamma^* = \arg \min_{\theta, \phi, \gamma} \ell_{cnt}(\mathcal{D}; \theta) + \ell_{snp}(\mathcal{D}; \theta, \phi) + \ell_{vid}(\mathcal{D}; \theta, \gamma) + \ell_{reg}(\mathcal{D}; \theta, \phi) \quad (9.1)$$

where $\ell_{cnt}(\cdot)$ denotes a contrastive loss that uses the mined hard and easy normal and abnormal snippet features, $\ell_{snp}(\cdot)$ is a loss function to train the snippet classifier $s_\phi(\cdot)$ using the top k snippet-level anomaly scores from normal and abnormal videos, $\ell_{vid}(\cdot)$ is a loss function to train the video classifier to predict whether the video contains anomalies, θ , ϕ and γ are respectively parameters of $\ell_{cnt}(\cdot)$, $\ell_{snp}(\cdot)$ and $\ell_{vid}(\cdot)$, and the regularisation loss is defined by

$$\ell_{reg}(\mathcal{D}; \theta, \phi) = \sum_{(\mathbf{F}_i, y_i) \in \mathcal{D}} \alpha \left(\frac{1}{T} \sum_{t=2}^T (\tilde{y}_i(t) - \tilde{y}_i(t-1))^2 \right) + \beta \left(\frac{1}{T} \sum_{t=1}^T |\tilde{y}_i(t)| \right), \quad (9.2)$$

with $\tilde{y}_i(t) \in [0, 1]$ denoting the anomaly classifier output for the t^{th} snippet from $\tilde{y}_i = s_\phi(f_\theta(\mathbf{F}_i))$. Note that in (11.8), the first term is a temporal smoothness regularisation, given that anomalous and normal events tend to be temporally consistent [211], the second term is the sparsity regularisation formulated based on the assumption that anomalous snippets are rare events in abnormal videos, and α and β are the hyper-parameters that weight both terms. Below, we describe the training of the video-level classifier, the snippet classifier, and the snippet contrastive loss.

Video Classifier Training. The video classifier is trained from a binary cross entropy loss to estimate if a video shows a polyp using the video-level labels. The loss $\ell_{vid}(\cdot)$ from (9.1) is the binary cross entropy loss defined as

$$\ell_{vid}(\mathcal{D}; \theta, \gamma) = - \sum_{(\mathbf{F}_i, y_i) \in \mathcal{D}} (y_i \log(v_\gamma(f_\theta(\mathbf{F}_i))) + (1 - y_i) \log(1 - v_\gamma(f_\theta(\mathbf{F}_i))))), \quad (9.3)$$

where $v_\gamma : \mathcal{X} \rightarrow [0, 1]$ is the video level anomaly classifier parameterised by γ .

Snippet Classifier Training. The snippet classifier is optimised by training a top k ranking loss function using a set that contains the k snippets with the largest anomaly scores from $s_\phi(\mathbf{F})$ in (9.1). More specifically, we propose the following loss $\ell_{snp}(\cdot)$ from (9.1) that maximises the separability between normal and abnormal videos:

$$\ell_{snp}(\mathcal{D}; \theta, \phi) = \sum_{\substack{(\mathbf{F}_i, y_i) \in \mathcal{D}, y_i=1 \\ (\mathbf{F}_j, y_j) \in \mathcal{D}, y_j=0}} \max(0, 1 - g_k(s_\phi(f_\theta(\mathbf{F}_i))) - g_k(s_\phi(f_\theta(\mathbf{F}_j))))), \quad (9.4)$$

where $g_k(\cdot)$ returns the mean anomaly score from $s_\phi(\cdot)$ of the top k snippets from a video [124, 217].

Contrastive Snippet Mining. To make anomaly classification robust to hard normal and abnormal snippets, we propose the following novel snippet contrastive loss:

$$\ell_{cnt}(\mathcal{D}; \theta) = \ell_c(\mathcal{D}^{HA}, \mathcal{D}^{EA}, \mathcal{D}^{EN}; \theta) + \ell_c(\mathcal{D}^{HN}, \mathcal{D}^{EN}, \mathcal{D}^{EA}; \theta), \quad (9.5)$$

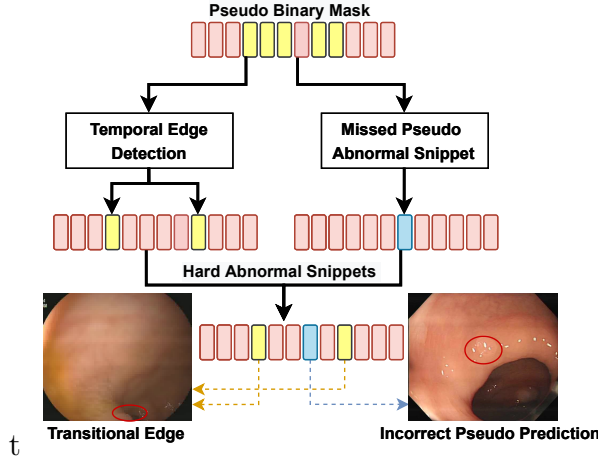


Figure 9.2: Hard abnormal snippet mining algorithm to select temporal edge snippets and missed pseudo abnormal snippets. Those two types of hard anomalies represent: 1) transitional frames where polyps may be partially visible, or 2) subtle (i.e., small and flat) polyps that can lead to incorrect low anomaly scores.

where \mathcal{D}^{HA} and \mathcal{D}^{EA} represent sets of hard and easy abnormal snippets, while \mathcal{D}^{HN} and \mathcal{D}^{EN} denote sets of hard and easy normal snippets,

$$\ell_c(\mathcal{D}^{HA}, \mathcal{D}^{EA}, \mathcal{D}^{EN}; \theta) = \sum_{\mathbf{F}_i \in \mathcal{D}^{HA}, \mathbf{F}_j \in \mathcal{D}^{EA}} \log \frac{\exp[\frac{1}{\tau} f_\theta(\mathbf{F}_i)^\top f_\theta(\mathbf{F}_j)]}{\exp[\frac{1}{\tau} f_\theta(\mathbf{F}_i)^\top f_\theta(\mathbf{F}_j)] + \sum_{\mathbf{F}_m \in \mathcal{D}^{EN}} \exp[\frac{1}{\tau} f_\theta(\mathbf{F}_i)^\top f_\theta(\mathbf{F}_m)]}, \quad (9.6)$$

and in a similar way we compute $\ell_c(\mathcal{D}^{HN}, \mathcal{D}^{EN}, \mathcal{D}^{EA}; \theta)$. The idea explored in (9.5) is to pull together easy and hard snippet features in \mathcal{X} from the same class (normal or abnormal) and push apart features from different classes.

The selection of \mathcal{D}^{HN} , \mathcal{D}^{EN} , \mathcal{D}^{HA} , \mathcal{D}^{EA} and their incorporation into our MIL learning framework is one key contribution of this work to address the poor detection accuracy of hard anomalous snippets in existing WVAD methods. Specifically, for abnormal videos, we first classify each of their T snippets with $\hat{y}(t) = (\tilde{y}(t) > \epsilon)$, where $\tilde{y} = s_\phi(f_\theta(\mathbf{F}))$. We then identify the temporal edge snippets and missed pseudo abnormal snippets as hard anomalies \mathcal{D}^{HA} . For temporal edge detection, we use the erosion operator to subtract the original and eroded sequences and locate such transitional edge snippets, which are considered as hard anomalies (See Fig. 9.2 - temporal edge detection), and inserted into \mathcal{D}^{HA} . For locating the missed pseudo abnormal snippets, we assume that a subtle anomalous event (i.e., a small/flat polyp) happens in a region of K consecutive snippets when $\frac{R}{K}$ (majority) of them have $\hat{y}(t) = 1$, where K and R are respectively the hyper-parameters to control the temporal length of the pseudo abnormal region and the ratio of the minimum number of the abnormal pseudo snippets inside that region. The incorrectly predicted normal snippets inside abnormal regions (i.e., missed abnormal snippets in Fig. 9.2) are also inserted into \mathcal{D}^{HA} as hard anomalies.

This hard anomaly selection process is motivated by the following two main observations: 1) subtle abnormal snippets from anomalous videos share similar characteristics to normal snippets (i.e., small and flat polyps) and consequently have low anomaly scores, and this can be easily identified from the adjacent abnormal snippets with higher anomaly scores since abnormal frames containing polyps are often contiguous; and 2) the transitional snippets between normal and abnormal events often contain noise such as water, endoscope pipe or partially visible polyps, so they are unreliable and can lead to inaccurate detection.

Hard normal (HN) snippets (e.g., healthy frames containing water and feces) are collected by selecting the snippets with top k anomaly scores from normal videos since normal videos do not have any abnormalities, so the ones with incorrectly predicted higher scores can be deemed as hard normal. For easy snippet mining, we hypothesise that the snippets with the smallest k anomaly scores from normal videos and the snippets with top k anomaly scores from abnormal videos are easy normal (EN) and easy abnormal (EA).

9.3 Experiments and Results

9.3.1 Dataset

To form a real-world large-scale video polyp detection dataset, we collected colonoscopy videos from two widely used public datasets: Hyper-Kvasir [21] and LDPolypVideo [147]. The new dataset contains 61 normal videos without polyps and 102 abnormal videos with polyps for training, and 30 normal videos and 60 abnormal videos for testing. The videos in the training set have video-level labels and the videos in testing set contain frame-level labels. This dataset contains over one million frames and has diverse polyps with various sizes and shapes, making it one of the largest and most challenging colonoscopy datasets in the field. The dataset setup will be publicly available upon paper acceptance.

9.3.2 Implementation Details

Following [211, 217], each video is divided into 32 video snippets, i.e., $T = 32$. For all experiments, we set $k = 3$ in (9.4). The 2048D input tokens are extracted from the 'mix_5c' layer of the pre-trained I3D [105] network. Note that the I3D network is not fine-tuned on any medical dataset. For the transformer block, we set the number of heads to 8, depth of transformer blocks to 12, and use a 3×1 DW Conv1D. α and β in (11.8) are both set to $5e - 4$. Our method is trained in an end-to-end manner using the Adam optimiser [107] with a weight decay of 0.0005 and a batch size of 32 for 200 epochs. The learning rate is set to 0.001. Following [211, 217], each mini-

Method	Publication	AUC	AP
DeepMIL [211]	CVPR'18	89.41	68.53
GCN-Ano [266]	CVPR'19	92.13	75.39
CLAWS [258]	ECCV'20	95.62	80.42
AR-Net [231]	ICME'20	88.59	71.58
MIST [73]	CVPR'21	94.53	72.85
RTFM [217]	ICCV'21	96.30	77.96
Ours		98.41	86.63

Table 9.1: Comparison of frame-level AUC and AP performance with other SOTA WVADs on colonoscopy dataset using the same I3D feature extractor.

batch consists of samples from 32 randomly selected normal and abnormal videos. The method is implemented in PyTorch [172] and trained with a NVIDIA 3090 GPU. The overall training times takes around 2.5 hours, and the mean inference time takes 0.06s per frame – this time includes the I3D extraction time. For all baselines, we use the same I3D backbone and benchmark setup as ours.

9.3.3 Evaluation on Polyp Frame Detection

Baselines. We train six SOTA WVAD baselines: DeepMIL [211], GCN-Ano [266], CLAWS [258], AR-Net [231], MIST [73], and RTFM [217]. The same experimental setup as our approach is applied to these baselines for fair comparison.

Evaluation Measures. Similarly to previous papers [56, 211], we use the frame-level area under the ROC curve (AUC) as the evaluation measure. Given that the AUC can produce optimistic results for imbalanced problems, such as anomaly detection, we follow [168, 245] and use average precision (AP) as another evaluation measure. Larger AUC and AP values indicate better performance.

Quantitative Comparison. We show the quantitative comparison results in Table 9.1. Our model achieves the best 98.4% AUC and 86.6% AP and outperforms all six SOTA methods by a large margin. We obtain a maximum 10% and a minimum 2% AUC improvement, and a maximum 18% and a minimum 6% AP improvement over the second best approaches. Our method substantially surpasses the most recent WVAD approach RTFM [217] by 8% AP.

Qualitative Comparison. In Fig. 9.3, we show the anomaly scores produced by our model for test videos from our polyp detection dataset. As illustrated by the orange curves, our model can effectively produce small anomaly scores for normal snippets and large anomaly scores for abnormal snippets. Our model is also able to detect multiple anomalous events (e.g., videos with two polyp event occurrences - first figure in Fig 9.3) in one video. Also, our model can also detect the subtle polyps (middle figure in Fig 9.3).

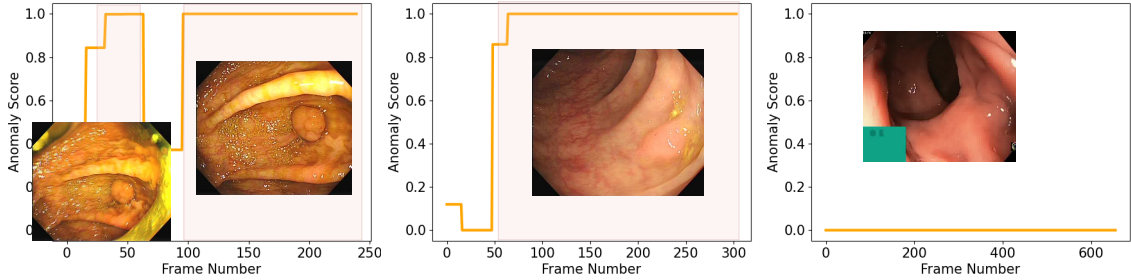


Figure 9.3: Anomaly scores (orange curve) of our method on test videos. Pink areas indicate the labelled testing abnormal events.

top-k (ℓ_{snp})	CTE	ℓ_{vid}	ℓ_{cnt}	AUC	AP
✓				92.88	71.96
✓	✓			94.92	79.56
✓	✓	✓		96.74	82.88
✓	✓	✓	✓	98.41	86.63

Table 9.2: Ablation studies for polyp frame detection. The linear network with top-k MIL ranking loss is considered as the baseline, and CTE denotes the Convolutional Transformer Encoder.

9.3.4 Ablation Study

Tab. 11.5 shows the contribution of each component of our proposed method on the testing set. The baseline top-k MIL network, trained with ℓ_{snp} , achieves 92.8% AUC and 71.9% AP. Our method obtains a significant performance gain by adding the proposed convolutional transformer encoder (CTE). Adding the video classifier, represented by the loss $\ell_{vid}(\cdot)$, boosts the performance by about 2% AUC and 3% AP. The proposed hard/easy snippet contrastive loss, denoted by the loss $\ell_{cnt}(\cdot)$, further improve the performance (e.g., increasing AP by about 4%), indicating the effectiveness of addressing the hard anomaly issues.

9.4 Conclusion

We proposed a new transformer-based MIL framework as a robust anomaly classifier for detecting polyp frames in colonoscopy videos. To the best of our knowledge, our method is the first to formulate polyp detection as a weakly-supervised video anomaly detection problem, and also to introduce transformer to explore global temporal dependency between video snippets. We also proposed a novel and effective contrastive snippet mining (CSM) to enable an effective learning of challenging abnormal polyp frames (i.e., small and partially visible polyps) and normal frames (i.e., water and fe-

ces). The resulting anomaly classifier showed SOTA results on our proposed large-scale colonoscopy dataset. Despite the remarkable performance on detecting polyp frames, our model may fail for online inference due to the transformer self-attention operation.

Statement of Authorship

Title of Paper	Few-Shot Anomaly Detection for Polyp Frames from Colonoscopy
Publication Status	<input checked="" type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	Published in International Conference on Medical Imaging Computing and Computer-Assisted Intervention (MICCAI) 2020.

Principal Author

Name of Principal Author (Candidate)	Yu Tian				
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper.				
Overall percentage (%)	90				
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.				
Signature	<table border="1" style="width: 100%;"> <tr> <td style="width: 80%;"></td> <td style="width: 20%;">Date</td> </tr> <tr> <td></td> <td>09/03/2022</td> </tr> </table>		Date		09/03/2022
	Date				
	09/03/2022				

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Gabriel Maicas				
Contribution to the Paper	Discussion and writing the revision.				
Signature	<table border="1" style="width: 100%;"> <tr> <td style="width: 80%;"></td> <td style="width: 20%;">Date</td> </tr> <tr> <td></td> <td>09/03/2022</td> </tr> </table>		Date		09/03/2022
	Date				
	09/03/2022				

Name of Co-Author	Leonardo Z.C.T. Pu				
Contribution to the Paper	Discussion and writing the revision.				
Signature	<table border="1" style="width: 100%;"> <tr> <td style="width: 80%;"></td> <td style="width: 20%;">Date</td> </tr> <tr> <td></td> <td>09/03/2022</td> </tr> </table>		Date		09/03/2022
	Date				
	09/03/2022				

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Rajvinder Singh		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Johan W Verjans		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Please cut and paste additional co-author panels here as required.

Chapter 10

Few-Shot Anomaly Detection for Polyp Frames from Colonoscopy

Abstract

Anomaly detection methods generally target the learning of a normal image distribution (i.e., inliers showing healthy cases) and during testing, samples relatively far from the learned distribution are classified as anomalies (i.e., outliers showing disease cases). These approaches tend to be sensitive to outliers that lie relatively close to inliers (e.g., a colonoscopy image with a small polyp). In this chapter, we address the inappropriate sensitivity to outliers by also learning from inliers. We propose a new few-shot anomaly detection method based on an encoder trained to maximise the mutual information between feature embeddings and normal images, followed by a few-shot score inference network, trained with a large set of inliers and a substantially smaller set of outliers. We evaluate our proposed method on the clinical problem of detecting frames containing polyps from colonoscopy video sequences, where the training set has 13350 normal images (i.e., without polyps) and less than 100 abnormal images (i.e., with polyps). The results of our proposed model on this data set reveal a state-of-the-art detection result, while the performance based on different number of anomaly samples is relatively stable after approximately 40 abnormal training images.

10.1 Introduction

Classification of rare events is a common problem in medical image analysis [131], e.g., disease detection in medical screening tests such as colonoscopy. In this scenario, normal images generally come from healthy patients, while abnormal images are from unhealthy ones, where the proportion of normal images in the training set tends to be substantially larger than the abnormal ones. One possible way to address such problems is through the design of training methods that can deal with imbalanced

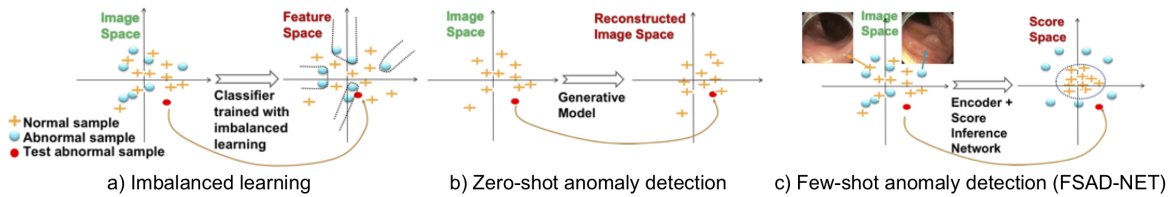


Figure 10.1: Depiction of the three different approaches to handle few-shot and zero-shot anomaly detection. Our proposed FSAD-NET demonstrate better deviations between normal and abnormal samples

learning problems [126, 128] (Fig. 10.1-(a)). Even though they are often effective, these approaches still need a fairly high number of abnormal training images. Alternatively, zero-shot anomaly detection methods [56, 139, 148, 273] tackle this problem using a training set containing only normal images to train a conditional generative model that can reconstruct normal images, and anomalies are detected based on the reconstruction errors of testing images (Fig. 10.1-(b)). Unfortunately, in practice these methods can misclassify outliers that lie relatively close to inliers (e.g., when cancer tissue occupies a small area of the image). Therefore, we propose a middle ground between these two approaches to address the issues of requiring a relatively large annotated data set and misclassifying challenging outliers.

In this chapter, we propose a few-shot anomaly detection method network (FSAD-NET) that is trained with a highly imbalanced training set, containing a large number of normal images (more than 10,000) and few abnormal images (less than 100) – Fig. 10.1-(c). The method first learns a feature encoder that is trained with normal images to maximise the mutual information (MI) between the training images and feature embeddings [98]. Next, we train a score inference network (SIN) [167] that pulls the feature embeddings of normal images close together toward a particular region of the feature space and pushes the embeddings of abnormal images away from that region of normal features.

In practice, FSAD-NET needs significantly less abnormal training images than typical imbalanced learning problems [126, 128]. Moreover, given that we access a few abnormal training images, FSAD-NET has the potential to be more effective at correctly classifying challenging outliers compared to typical zero-shot anomaly detection methods [56, 139, 148, 273]. To the best of our knowledge, our method is the first medical image analysis work to explore few-shot anomaly detection with a feature encoder that maximises MI between training images and embeddings, and explicitly optimises anomaly scores. We evaluate FSAD-NET on the detection of colonoscopy video frames that contain polyps with a training set of more than 10000 normal images (without polyps) and less than 100 abnormal images. Results show that our FSAD-NET is more accurate than previous zero-shot anomaly detection approaches, which allows us

to conclude that incorporating few abnormal cases into the training process improves the performance of anomaly detection methods. Our approach also shows better accuracy than imbalanced learning methods, suggesting that FSAD-NET is more effective at dealing with very small training sets of abnormal images.

10.2 Related Work

Colorectal cancer is considered to be one of the most harmful cancers [67, 181]. One effective method for screening patients for colorectal cancer is colonoscopy, where the goal is to detect polyps that are malignant or pre-malignant using a camera that is inserted into the bowel. Accurate early detection of polyps may improve the 5-year survival rate to over 90% [62]. Unfortunately, the accuracy and speed of manual polyp detection can be affected by human factors, such as fatigue and expertise [180, 181, 225]. Therefore, automated polyp detection systems could help doctors improve polyp detection accuracy during a colonoscopy [67, 181]. Traditional systems to detect polyps are based on a supervised two-class classifier [67, 111] trained with large training sets of images without polyps (i.e. normal) and images containing polyps (i.e. abnormal). Annotation of such training sets is unfortunately difficult because the vast majority of colonoscopy video frames contain normal images, making the manual search for images that contain polyps challenging. Imbalanced learning solutions can therefore be used in this context [126, 128], but its extension to polyp detection may not be effective without a relatively large number of abnormal images in the training set. Because of this limitation, zero-shot anomaly detection methods have been studied [60, 64, 65, 139, 167, 174, 199], where the idea is to learn a distribution of normal images in a particular feature space, to subsequently test samples that do not fit well in this distribution and are then classified as an outlier that may contain a polyp.

Zero-shot anomaly detection methods assume that the conditional generative model [56, 64, 139, 148, 174, 199, 273]) can only reconstruct normal data. Hence, when presented with an abnormal test image, the model produces a large reconstruction error. However, using an image reconstruction error for training is an indirect optimisation of the anomaly score, which can lead to a sub-optimal training process. For example, an abnormal image with a small polyp may have a low reconstruction error because the small area affected by the polyp and can be wrongly classified as normal. We advocate that the performance of zero-shot anomaly detection methods can improve with the use of a small set of abnormal training images (less than 100). Such imbalance learning problem has been tackled by few-shot classification approaches before. However, our problem has a different setup compared to problems handled by traditional few-shot learning methods that generally have many few-shot balanced multi-class problems for training [74, 160, 213], while ours has only one few-shot highly imbalanced binary problem for training. Hence, we can only compare our method with baseline approaches that

handle imbalance learning [128, 188]. For instance, Ren et al. [188] propose a learning algorithm for highly imbalanced learning problems that weights training samples using a balanced validation set – the need for this validation set is a disadvantage of this approach. The focal loss approach [128] is effective at handling imbalanced learning, but it may still need a large number of samples from both classes.

Few-shot anomaly detection has been shown in a non-medical image analysis context with the method SIN [164] that is designed to directly optimise an anomaly score for normal and abnormal images. The main challenge to train SIN lies in the high dimensionality of the images [164]. Therefore, one way to alleviate this challenge is to introduce a dimensionality reduction before training SIN. Recently, deep infomax (DIM) [98] has been shown to be an effective dimensionality reduction approach. In our paper, we propose a method that uses DIM to learn a low-dimensionality feature embedding that is then used by SIN to classify anomalies.

10.3 Data Set and Method

10.3.1 Dataset

The data set is obtained from 18 colonoscopy videos from 15 patients. Video frames containing blurred visual information are removed using the variance of Laplacian method [66]. We then sub-sample consecutive frames by taking one of every five frames because the correlation between them makes training ineffective. We also remove frames containing feces and water to reduce the need for a very large normal training set (we plan to handle such distractors in future work). This data set is defined by $\mathcal{D} = \{(\mathbf{x}, d, y)_i\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x} : \Omega \rightarrow \mathbb{R}^3$ denotes a colonoscopy frame (Ω represents the frame lattice), $d \in \mathbb{N}$ represents patient identification¹, $y \in \mathcal{Y} = \{Normal, Abnormal\}$ denotes the normal (without polyp) and abnormal (with polyp) classes. The distribution of this data set is as follows: 1) Training set: a set of 13250 normal images (without polyps), denoted by $\mathcal{D}_N \subset \mathcal{D}$, and a set containing between 10 and 80 abnormal images, denoted by $\mathcal{D}_A \subset \mathcal{D}$; 2) Validation set: 100 normal images and 100 abnormal images for model selection; and 3) Testing set: 967 images, with 217 (25% of the set) abnormal images and 700 (75% of the set) normal images. The patients in the testing set do not appear in the training/validation sets and vice versa. This abnormality proportion (on the testing set) is commonly defined in other anomaly detection literature [174, 199]. These frames were obtained with the Olympus (®)190 dual focus endoscope.

¹Note that the data set has been de-identified, so d is useful only for splitting \mathcal{D} into training, testing and validation sets in a patient-wise manner.

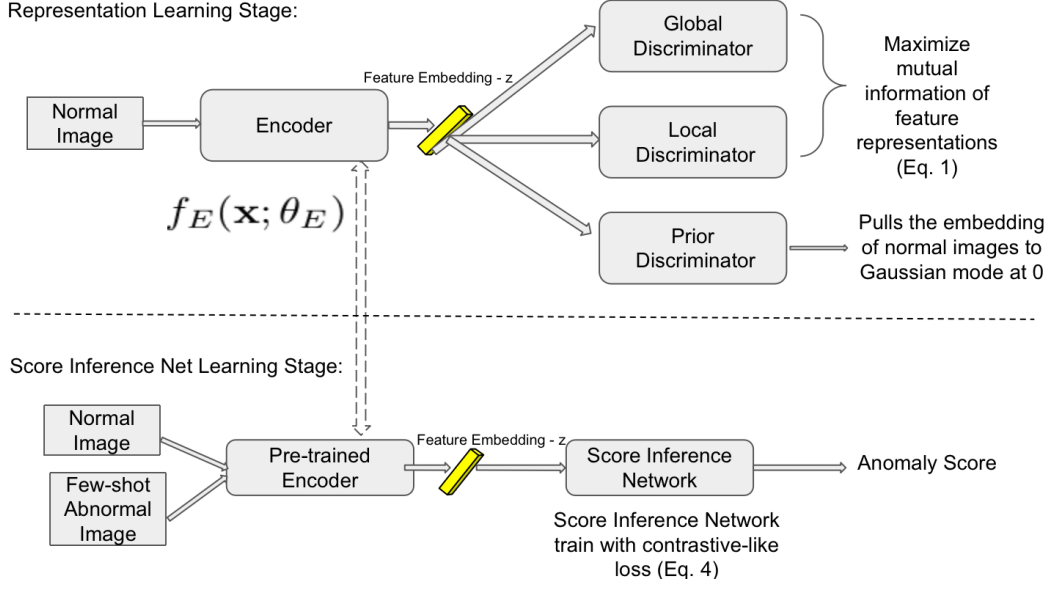


Figure 10.2: The first stage of FSAD-NET training consists of modelling the encoder by maximising the MI between normal training images and embeddings in a global and local manner and by minimising the divergence of embeddings and a prior distribution [98]. The embeddings produced by the encoder are then used to train the SIN using a contrastive-like loss [167].

10.3.2 Method

The training process of our proposed FSAD-NET (Fig. 10.3) is divided into two stages: 1) pre-training of a feature encoder $\mathbf{z} = f_E(\mathbf{x}; \theta_E)$ (θ_E is the encoder parameter and $\mathbf{z} \in \mathbb{R}^Z$) to learn an image embedding that maximises the mutual information (MI) between normal images $\mathbf{x} \in \mathcal{D}_N$ and their embeddings \mathbf{z} [98]; and 2) training of the SIN $f_S(f_E(\mathbf{x}; \theta_E); \theta_S)$ [167], parameterised by θ_S , with a contrastive-like loss that uses \mathcal{D}_N and \mathcal{D}_A to achieve the goal $f_S(f_E(\mathbf{x} \in \mathcal{D}_A; \theta_E); \theta_S) > f_S(f_E(\mathbf{x} \in \mathcal{D}_N; \theta_E); \theta_S)$.

More specifically, the training of the encoder to maximise the MI between the normal samples $\mathbf{x} \in \mathcal{D}_N$ and their feature embeddings $\mathbf{z} = f_E(\mathbf{x} \in \mathcal{D}_N; \theta_E)$ [98] is achieved with

$$\begin{aligned} \theta_E^*, \theta_G^*, \theta_L^* = \arg \max_{\theta_E, \theta_G, \theta_L} & \left(\alpha \hat{I}_{\theta_G}(\mathbf{x}; f_E(\mathbf{x}; \theta_E)) + \frac{\beta}{|\mathcal{M}|} \sum_{\omega \in \mathcal{M}} \hat{I}_{\theta_L}(\mathbf{x}(\omega); f_E(\mathbf{x}(\omega); \theta_E)) \right) \\ & + \gamma \arg \min_{\theta_E} \arg \max_{\phi} \hat{D}_{\phi}(\mathbb{V} || \mathbb{U}_{\mathbb{P}, \theta_E}) \end{aligned} \quad (10.1)$$

where α, β, γ are the model hyperparameters, the functions $\hat{I}_G(\cdot)$ and $\hat{I}_L(\cdot)$ denote an

MI lower bound based on the Donsker-Varadhan representation of the Kullback-Leibler (KL)-divergence [98], defined by

$$\hat{I}_{\theta_G}(\mathbf{x}; f_E(\mathbf{x}; \theta_E)) = \mathbb{E}_{\mathbb{J}}[f_G(\mathbf{x}, f_E(\mathbf{x}; \theta_E); \theta_G)] - \log \mathbb{E}_{\mathbb{M}}[e^{f_G(\mathbf{x}, f_E(\mathbf{x}; \theta_E); \theta_G)}], \quad (10.2)$$

with \mathbb{J} denoting the joint distribution between images \mathbf{x} and their respective embeddings $\mathbf{z} = f_E(\mathbf{x}; \theta_E)$, \mathbb{M} representing the product of the marginals of the images and embeddings, and $f_G(\mathbf{x}, f_E(\mathbf{x}; \theta_E); \theta_G)$ being a discriminator parameterised by θ_G . Also in (10.1), the function $\hat{I}_{\theta_L}(\mathbf{x}(i); f_E(\mathbf{x}(i); \theta_E))$, defined similarly as (10.2) for the discriminator $f_L(\mathbf{x}(\omega), f_E(\mathbf{x}(\omega); \theta_E); \theta_L)$, is the local MI between image regions $\mathbf{x}(\omega)$ ($\omega \in \mathcal{M} \subset \Omega$) and respective local embeddings $f_E(\mathbf{x}(\omega), \theta_E)$. Moreover in (10.1),

$$\arg \min_{\theta_E} \arg \max_{\phi} \hat{D}_{\phi}(\mathbb{V} || \mathbb{U}_{\mathbb{P}, \theta_E}) = \mathbb{E}_{\mathbb{V}}[\log d(\mathbf{z}; \phi)] + \mathbb{E}_{\mathbb{P}}[\log(1 - d(f_E(\mathbf{x}; \theta_E); \phi))], \quad (10.3)$$

with \mathbb{V} denoting a prior distribution for the embeddings \mathbf{z} (\mathbb{V} is assumed to be a normal distribution $\mathcal{N}(\cdot; \mu_{\mathbb{V}}, \Sigma_{\mathbb{V}})$, with mean $\mu_{\mathbb{V}}$ and covariance $\Sigma_{\mathbb{V}}$), \mathbb{P} the distribution of the embeddings $\mathbf{z} = f_E(\mathbf{x} \in \mathcal{N}_N; \theta_E)$, and $d(\cdot; \phi)$ is a discriminator modelled with adversarial training to estimate the likelihood that the input is sampled from \mathbb{V} or \mathbb{P} . This objective function pulls the feature embeddings of the normal images toward $\mathcal{N}(\cdot; \mu_{\mathbb{V}}, \Sigma_{\mathbb{V}})$.

The next step of the learning process consists of computing the embeddings of normal and abnormal images with $\mathbf{z} = f_E(\mathbf{x} \in \mathcal{D}_A \cup \mathcal{D}_N; \theta_E^*)$ to train $f_S(\mathbf{z}; \theta_S)$ using a contrastive-like loss to directly optimise the anomaly score [167]. More specifically, the contrastive loss for each training sample is defined as:

$$\ell_S = \mathbb{I}(y \text{ is Normal}) |s(f_S(\mathbf{z}; \theta_S))| + \mathbb{I}(y \text{ is Abnormal}) \max(0, a - s(f_S(\mathbf{z}; \theta_S))), \quad (10.4)$$

where $\mathbb{I}(\cdot)$ is an indicator function that is equal to one when the condition in the parameter is true, and zero otherwise, $s(x) = \frac{x - \mu_S}{\sigma_S}$ with $\mu_S = 0$ and $\sigma_S = 1$ representing the mean and standard deviation of the prior distribution for the anomaly scores for normal images, and a is the minimum margin between μ_S and the anomaly scores of abnormal images [167]. The loss in (10.4) pulls the scores from normal images to μ_S and pushes the scores of abnormal images away from μ_S with a margin of at least a .

During inference, we take a test image \mathbf{x} , compute the feature embedding with $f_E(\mathbf{x}; \theta_E)$ and then compute the score with $s = f_S(\mathbf{z}; \theta_S)$ – the score result s is then compared to a threshold τ to determine if the test image is normal or abnormal. We considered the score s as the estimation of the notion of closeness which is related to the likelihood that the embedding of a colonoscopy image is classified as belonging to the set of normal images.

10.4 Experiment

10.4.1 Experimental Setup

The original colonoscopy images are resized from initial resolution $1072 \times 1072 \times 3$ to $64 \times 64 \times 3$ to reduce the training and inference computational costs. We found that $64 \times 64 \times 3$ is the minimum size that we can use without a negative impact on AUC. We note the polyps are still visible at such resolution, as shown in Fig 10.4. The model selection (to select optimiser, learning rate and model structure) is done using the validation set mentioned in Sec. 10.3.1. We use Adam [57] optimiser during training with a learning rate of 0.0001 for the encoder and SIN learning. We adopt batch normalisation for both stages. We make sure our method uses a similar backbone architecture as other competing approaches in Tab. 10.1. In particular, the encoder $f_E(\cdot; \theta_E)$ uses four convolution layers (with 64, 128, 256, 512 filters of size 4×4). The global discriminator $f_G(\cdot; \theta_G)$ has three convolutional layers (with 128, 64, 32 filters of size 3×3). The local discriminator $f_L(\cdot; \theta_L)$ has three convolutional layers (with 192, 512, 512 filters of size 1×1). The prior discriminator $d(\cdot; \phi)$ has three linear layers with 1000, 200, 1 nodes per layer). We also use the validation set to estimate $a = 6$ in (10.4). In (10.1), we follow the DIM paper for setting the hyper-parameters as follows [98]: $\alpha = 0.5$, $\gamma = 1$, $\beta = 0.1$. For the prior distribution for the embeddings in (10.3), we set $\mu_V = \mathbf{0}$ (i.e., a Z -dimensional vector of zeros), and Σ_V is a $Z \times Z$ identity matrix. To train the model, we first train the encoder, local, global and prior discriminator (representation learning stage) for 6000 epochs with a mini-batch of 64 samples. We then train SIN for 1000 epochs, with a batch size of 64, while fixing the parameters of encoder, local, global and prior discriminator. We implement our method using Pytorch [171]. The detection results are measured with the area under the receiver operating characteristic curve (AUC) on the test set [174, 199], computed by varying the inference threshold τ for the score result s .

10.4.2 Anomaly Detection Results

The test set AUC results shown in Table 10.1 are divided into zero-shot and few-shot. The zero-shot rows show results obtained from the following zero-shot anomaly detection methods²: ADGAN [139], OCGAN [174], f-anogan and its variants [199] that involve image-to-image mean square error (MSE) loss (izi), Z-to-Z MSE loss (ziz) and its hybrid version (izif). Our FSAD-NET model outperforms all zero-shot learning methods by a large margin, showing the importance of using a few abnormal samples for training. For the few-shot results, we consider the cases where we have 30 and 40 abnormal training images, and we test several variants of the FSAD-NET. We use

²Codes were downloaded from the authors' Github pages and tuned for our problem.

Table 10.1: Comparison between our proposed FSAD-NET and other state of the art zero-shot and few-shot anomaly detection methods.

	Methods	AUC
Zero-Shot	DAE [60]	0.6384
	VAE [51]	0.6528
	OC-GAN [174]	0.6137
	f-AnoGAN(ziz) [199]	0.6629
	f-AnoGAN(izi) [199]	0.6831
	f-AnoGAN(izif) [199]	0.6997
	ADGAN [139]	0.7391
Few-Shot	Densenet121 [99] (40 abnormal samples)	0.8231
	cross-entropy (30 abnormal samples)	0.6826
	cross-entropy (40 abnormal samples)	0.7115
	Focal loss (30 abnormal samples)	0.7038
	Focal loss (40 abnormal samples)	0.7235
	without RL (40 abnormal samples)	0.6011
	Learning to Reweight [188] (40 abnormal samples)	0.7862
	AE network (30 abnormal samples)	0.819
	AE network (40 abnormal samples)	0.835
	FSAD-NET (30 abnormal samples)	0.855
	FSAD-NET (40 abnormal samples)	0.9033

between 30 and 40 abnormal training images because that is the range, where we observe that our FSAD-NET produces stable AUC results. As a baseline approach, we train Densenet121 [99] using high levels of data augmentation to deal with the training imbalance issue. However, our FSAD-Net outperforms Densenet121 by a large margin. The variants of FSAD-NET are designed to test the importance of each stage of our method. The methods labelled as 'Cross entropy' and 'Focal loss' replace the contrastive loss in (10.4) by the cross entropy loss (commonly used in classification problems) [78] and the focal loss (robust to imbalanced learning problems) [128], respectively. FSAD-NET shows substantially better results, indicating the importance of using a more appropriate loss function for few-shot anomaly detection. To show the importance of representation learning (RL) in FSAD-Net, we tested FSAD-Net without it, which shows much lower AUC results than competing approaches. Also, we compared our method with a few-shot learning baseline [188], which proposes a learning algorithm for highly imbalanced learning problems. When used to train FSAD-Net, it achieved 78.62% of mean AUC when training with 40 abnormal training samples. Hence our model shows more accurate results than that approach. Furthermore, we test the importance of DIM to train the encoder in (10.1) by replacing it by the deep

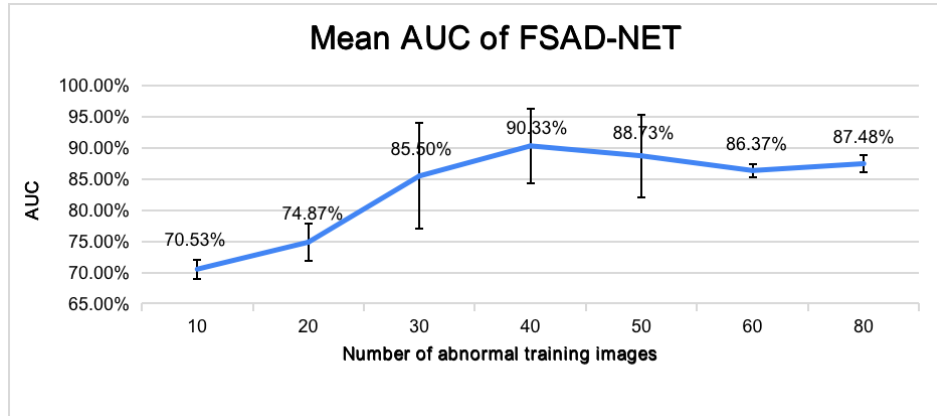


Figure 10.3: AUC mean and standard deviation of FSAD-NET computed over different number of abnormal training images.

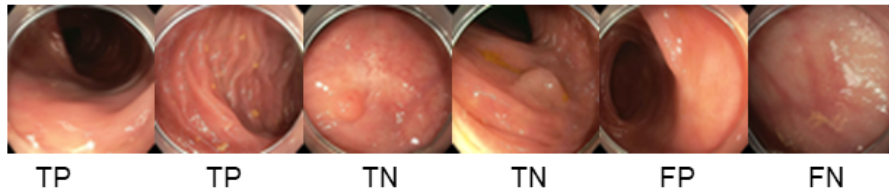


Figure 10.4: True positive (TP), true negative (TN), false positive (FP) and false negative (FN) results produce by FSAD-NET (Negative = frame with polyp).

auto-encoder [60] (labelled as AE network) – results show that FSAD-NET is more accurate, indicating the effectiveness of using MI and prior distribution for learning the feature embeddings in (10.1).

We further investigate the performance of our proposed FSAD-NET as a function of the number of abnormal training images that can vary from 10 to 80. For each number of abnormal training images, we train our model three times, using different training sets each time, and we compute the mean and standard deviation of the AUC results. The result of this experiment in Fig. 10.3 shows that: 1) the performance stabilises between 85%-90% when feeding the model 30 or more abnormal training images; and 2) our method is robust to extremely small training sets of abnormal images. We show a few true positive, true negative, false positive and false negative results produce by FSAD-NET in Fig. 10.4.


10.5 Conclusion

We propose the first few-shot anomaly detection framework, named as FSAD-NET, for medical image analysis applications. FSAD-NET consists of an encoder trained to maximise the mutual information between normal images and respective embeddings and a score inference network that classifies between normal and abnormal colonoscopy frames. Results show that our method achieves state-of-the-art anomaly detection performance on our colonoscopy data set, compared to previous zero-shot anomaly detection methods and imbalanced learning methods. In the future, we expect to extend our approach to polyp localisation and to work with colonoscopy frames containing distractors, like feces and water.

Statement of Authorship

Title of Paper	Pixel-wise Energy-biased Abstention Learning for Anomaly Segmentation on Complex Urban Driving Scenes
Publication Status	<input type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input checked="" type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	Submitted to ECCV 2022


Principal Author


Name of Principal Author (Candidate)	Yu Tian
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper (Equal contribution with Yuyuan Liu).
Overall percentage (%)	40
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	
Date	09/03/2022

Co-Author Contributions

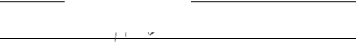
By signing the Statement of Authorship, each author certifies that:

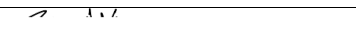
- i. the candidate's stated contribution to the publication is accurate (as detailed above);
- ii. permission is granted for the candidate to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.


Name of Co-Author	Yuyuan Liu
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper.
Signature	
Date	09/03/2022

Name of Co-Author	Guansong Pang
Contribution to the Paper	Discussion and writing the revision.
Signature	
Date	09/03/2022

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Fengbei Liu		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Yuanhong Chen		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	Discussion and writing the revision.		
Signature		Date	09/03/2022

Please cut and paste additional co-author panels here as required.

Chapter 11

Pixel-wise Energy-biased Abstention Learning for Anomaly Segmentation on Complex Urban Driving Scenes

Abstract

State-of-the-art (SOTA) anomaly segmentation approaches on complex urban driving scenes explore pixel-wise classification uncertainty learned from outlier exposure, or external reconstruction models. However, previous uncertainty approaches that directly associate high uncertainty to anomaly may sometimes lead to incorrect anomaly predictions, and external reconstruction models tend to be too inefficient for real-time self-driving embedded systems. In this chapter, we propose a new anomaly segmentation method, named pixel-wise energy-biased abstention learning (PEBAL), that explores pixel-wise abstention learning (AL) with a model that learns an adaptive pixel-level anomaly class, and an energy-based model (EBM) that learns inlier pixel distribution. More specifically, PEBAL is based on a non-trivial joint training of EBM and AL, where EBM is trained to output high-energy for anomaly pixels (from outlier exposure) and AL is trained such that these high-energy pixels receive adaptive low penalty for being included to the anomaly class. We extensively evaluate PEBAL against the SOTA and show that it achieves the best performance across four benchmarks.

11.1 Introduction

Recent advances in semantic segmentation have shown tremendous improvements on complex urban driving scenes [116]. Despite the accurate predictions on the inlier classes, the model fails to properly recognise anomalous objects that deviate from

the training inlier distribution (col. 2 of Fig. 11.1). Addressing such failure cases is crucial to road safety for autonomous driving vehicles. For example, anomalies can be represented by unexpected objects in the middle of the road, such as a large rock or an unexpected animal that can be incorrectly predicted as a part of the road class, leading to potentially fatal traffic collisions.

Current methods [17, 20, 27, 49, 103, 130, 155, 246] to detect and segment anomalous objects in complex urban driving scenes tend to depend on classification uncertainty or image reconstruction. The association of high classification uncertainty with anomaly is intuitive, but it has a few caveats. For instance, classification uncertainty happens when samples are close to classification decision boundaries, but there is no guarantee that all anomalies will be close to classification boundaries. Furthermore, samples close to classification boundaries may not be anomalies at all, but just hard inlier samples. Hence, these uncertainty based methods may detect a large number of false positive and false negative anomalies. For example, Fig. 11.1 shows that the previous SOTA Meta-OoD [27] misses important anomalous pixels (all rows), while misclassifying anomalies (e.g., vegetation in rows 1, 2, 3), even with the use of the outlier exposure (OE) strategy [96]. In fact, the OE strategy maximises the uncertainty for proxy anomalies, which can cause the model to be more uncertain for all inlier classes and detect false positive anomalies (e.g., Meta-OoD mis-classifies trees or bush with high anomaly scores – Fig. 11.1 col 4). Reconstruction methods [49, 246] add an extra network to reconstruct the input images from the estimated segmentation, where differences are assumed to be anomalous. Not only does this approach depend on accurate segmentation results for precise reconstruction, but they also require an extra reconstruction network that is hard to train and inefficient to run in real-time self-driving embedded systems. Moreover, reconstruction methods that rely on a discrepancy module require re-training whenever the inlier segmentation model changes due to input distribution shift [49], limiting their applicability in real-world systems. Furthermore, previous approaches [17, 27, 49, 82, 103, 130] ignore a couple of important constraints for anomaly segmentation, namely smoothness (e.g., Meta-OoD fails to classify neighbouring anomaly pixels in Fig. 11.1, rows 1, 4) and sparsity (e.g., Meta-OoD incorrectly detects a large number of anomalous pixels—see yellow and red regions in Fig. 11.1, rows 1, 2, 3). Another common issue shared by previous methods [17, 27, 130] is that they usually rely on the re-training of the entire network for OE, which is inefficient and can also bias the classification towards outliers.

In this chapter, we propose a new anomaly segmentation method, the pixel-wise energy-biased abstention learning (PEBAL), that directly learns a pixel-level anomaly class, in addition to the pre-defined inlier classes, to reject/abstain anomalous pixels that are dissimilar to any of the inlier classes. It is achieved by a joint optimisation of a novel pixel-wise anomaly abstention learning (PAL) and an energy based model (EBM) [81, 117, 138]. Particularly, abstention learning (AL) [142] was originally de-

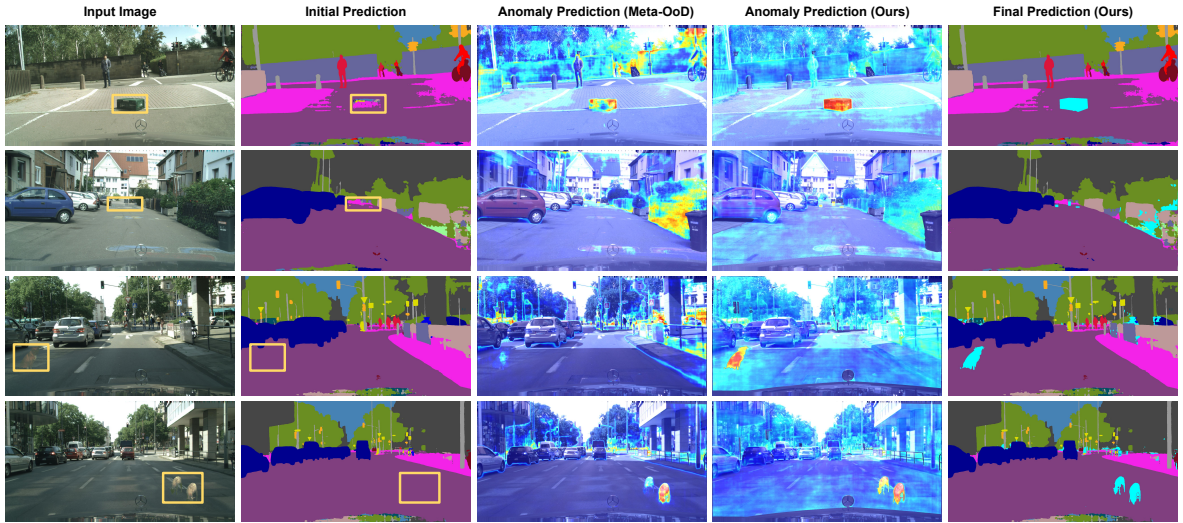


Figure 11.1: **Anomaly segmentation overview.** From the **input image** (anomaly highlighted with a yellow box), the **initial prediction** shows the original segmentation results with anomalies classified as a one of the pre-defined inlier classes. **Anomaly predictions** by the previous SOTA **Meta-OoD** [27] and **our** method show an anomaly map with high scores (in yellow and red) for anomalous pixels, where our approach shows less false positive and false negative detections. Consequently, our method can detect small and distant anomalies (row 2) and blurry/unclear anomalies (rows 1, 3, 4) more accurately than Meta-OoD [27]. In our **final prediction**, anomalous pixels are coloured as cyan. **Some anomalies are small and blurred (e.g., row 2), so please zoom in the PDF for better visualisation.**

veloped to learn an image-level anomaly class, which is significantly challenged by the pixel-wise anomaly segmentation task that requires pixel-level anomaly class learning. This is because the original AL model treats all pixel inputs equally with a single pre-defined fixed penalty factor to regularise the classification of anomalous pixels, while adaptive penalties are typically required for different pixels in a complex driving scene, e.g., pixels in small (distant) objects vs. large (near) objects, or centred pixels vs fringe pixels of objects. PEBAL is designed to address this issue by learning adaptive pixel-wise energy-based penalties, which automatically decreases the penalty for pixels that are likely to be anomalies. Hence, our model does not explore previously proposed uncertainty measures (e.g., entropy or softmax criteria) or image reconstruction, and instead, for the first time, explicitly learns a new pixel-wise anomaly class. The learned penalty factors are jointly optimised with EBM, resulting in a mutually beneficial optimisation of anomaly and inlier segmentation. Additionally, we impose smoothness and sparsity constraints to the learning of the anomaly segmentation by PEBAL, incorporating local and global dependencies into the pixel-wise penalty esti-

mation and anomaly score learning. Finally, the training of PEBAL is efficient given that we only need to fine-tune the last block of the segmentation model to achieve accurate inference. To summarise, our contributions are the following:

- We propose the pixel-wise energy-biased abstention learning (PEBAL) that jointly optimises a novel pixel-wise anomaly abstention learning (PAL) and energy-based models (EBM) to learn adaptive pixel-level anomalies. PEBAL mutually reinforces PAL and EBM in detecting anomalies, enabling accurate segmentation of anomalous pixels without compromising the segmentation of inlier pixels (cols. 4,5 of Fig. 11.1).
- We introduce a new pixel-wise energy-biased penalty estimation, which can learn adaptive energy-based penalties to highly varying pixels in a complex driving scene, allowing a robust detection of small/distant and blurry anomalous objects (Fig. 11.1 row 2).
- We further refine our PEBAL training, using a novel smoothness and sparsity regularisation on anomaly scores to consider the local and global dependencies of the pixels, enabling the reduction of false positive/negative anomaly predictions.

We validate our approach on Fishyscapes leaderboard [20], and achieve SOTA classification accuracy on all relevant benchmarks. We also achieve the best classification results on LostandFound [179] and Road Anomaly [130] test sets, significantly surpassing other competing methods. We also show that our approach produces competitive pixel-wise calibration results on Cityscapes [39].

11.2 Related work

Uncertainty-based Anomaly Segmentation. Early uncertainty-based methods [95, 118, 127] focused on the estimation of image-level anomalies, but they tended to misclassify object boundaries as anomalies [103]. Jung et al. [103] mitigate this issue by iteratively replacing false anomalous boundary pixels with neighbouring non-boundary pixels that have low anomaly score. In [106, 115, 155], the boundary issue was tackled with a pixel-wise uncertainty estimated with MC dropout, but they showed a low pixel-wise anomaly detection accuracy [130]. Without fine-tuning using a proxy outlier dataset, uncertainty estimation may not be accurate enough to detect anomalies and can predict high uncertainty for challenging inliers or low uncertainty for outliers due to overconfident misclassification.

Reconstruction-based Anomaly Segmentation. Anomalies can also be segmented from the errors between the input image and its reconstruction obtained from its predicted segmentation map [8, 33, 40, 49, 86, 130, 230, 246]. Those approaches are challenged by the dependence on an accurate segmentation prediction, by the complexity

of reconstruction models that usually require long training and inference processes, and also by the low quality of the reconstructed images.

Anomaly Segmentation via Outlier Exposure. Hendrycks et al. [96] propose the outlier exposure (OE) strategy that uses an auxiliary dataset of outliers that do not overlap with the real outliers/anomalies to improve the anomaly detection performance. This OE strategy uses outliers from ImageNet [17, 18, 226], void class of Cityscape [49] or COCO [27], where the expectation is that the model can generalise to unseen outliers. Maximising uncertainty for outliers using the OE strategy can lead to a deterioration of the segmentation of inliers [18, 226]. Another major drawback of OE methods is that they are trained using outlier images or objects without considering the fact that outliers are rare events that appear around inliers. Hence, the training contains a disproportionately high amount of outliers [27] that can bias the segmentation toward the anomaly class. We address this issue by respecting the anomaly detection assumption, where anomalous objects are rare, contribute to a small proportion of the training set, and appear around inliers.

Abstention Learning. The abstention learning mechanism [54] adds a “reserve” (i.e., anomaly) class that is predicted when the classification predictions for all inlier classes are not high enough. This method shows good performance in learning holistic image-level anomaly class with a single pre-defined penalty factor for the whole training set, but it fails to learn fine-grained pixel-level anomaly class as an adaptive pixel-wise penalty is required for highly varying pixel-level anomalies (see Table 11.5). We address this issue by learning a novel pixel-wise energy-biased penalty estimator that is jointly trained with fine-grained abstention learning. It is worth noting that differently from uncertainty-based methods [20, 27, 93, 103] that assume anomaly even when the model is uncertain but confident, abstention learning requires all classes to have low confidence to predict the anomaly class.

Energy-based Models. EBM is trained such that inlier training samples have low energy, whereas non-training outlier samples (i.e., anomalies) are expected to have high energy [117]. This energy value can then be used to compute the probability of a sample to belong to the inlier distribution. Recently, EBMs are being implemented with deep learning models [81, 138, 161], and to learn them, it is necessary to compute the partition function, which is generally estimated with Markov Chain Monte Carlo (MCMC) [81], but this estimation cannot generate accurate high-resolution images. Hence, we follow the simpler idea of estimating the energy score with the *logsumexp* operator [81, 138], where we minimise the energy of inliers and use an OE strategy [96] to maximise the energy of outliers. Hence, we do not need to compute the partition function.

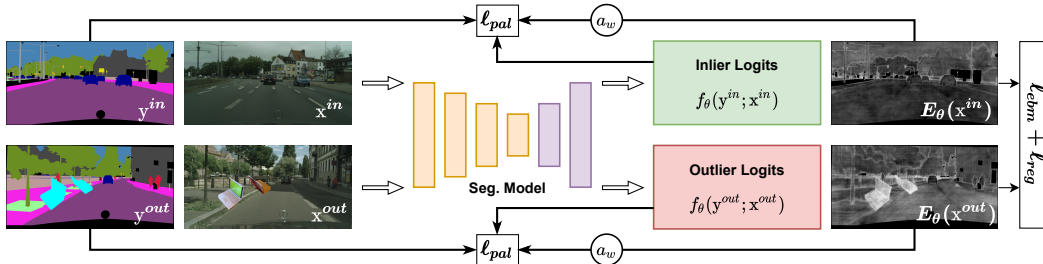


Figure 11.2: **PEBAL**. The pixel-wise anomaly abstention (PAL) loss ℓ_{pal} learns to abstain the prediction of outlier pixels from \mathbf{x}^{out} containing OE objects (i.e., cyan coloured masks) and calibrate the logit of inlier classes (i.e., reduction of the inlier logits) from both inlier image \mathbf{x}^{in} and outlier image \mathbf{x}^{out} . The EBM loss ℓ_{ebm} pushes the free energy E_θ to low values for inlier pixels and pulls that to high values for outlier pixels, where a regularisation loss ℓ_{reg} enforces the smoothness and sparsity constraints on the energy maps. Such EBM learning reduces the logit of inlier classes to share similar values at the same time, facilitating the ℓ_{pal} learning. Then, the pixel-wise penalty a_ω associated with the abstention class at position ω is estimated to bias the penalty to be low for outlier pixels and high for inlier pixels, which in turn encourages high free energy for anomalies and enforces ℓ_{pal} to abstain the anomalous pixels.

11.3 Method

We present our PEBAL in this section (see Fig. 11.2), where we first describe the dataset, then introduce abstention learning and EBM. Next, we present the loss function to train the model, followed by the training and inference procedures.

11.3.1 Training Set

We assume to have a set of inlier training images and annotations $\mathcal{D}^{in} = \{(\mathbf{x}_i, \mathbf{y}_i^{in})\}_{i=1}^{|\mathcal{D}^{in}|}$, where $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{H \times W \times C}$ denotes an image with C colour channels, and $\mathbf{y}^{in} \in \mathcal{Y}^{in} \subset \{0, 1\}^{H \times W \times Y}$ denotes the inlier pixel level labels that can belong to Y classes. We also have a set of outlier images and annotations $\mathcal{D}^{out} = \{(\mathbf{x}_i, \mathbf{y}_i^{out})\}_{i=1}^{|\mathcal{D}^{out}|}$, where $\mathbf{y}^{out} \in \mathcal{Y}^{out} \subset \{0, 1\}^{H \times W \times (Y+1)}$ denotes the outlier pixel-level labels, with the class $Y + 1$ reserved for pixels belonging to the anomaly class. Note that similarly to previous papers [27], the types of anomalies in training set \mathcal{D}^{out} do not overlap with the anomalies to be found in the testing set.

11.3.2 Pixel-wise Energy-biased Abstention Learning (PEBAL)

The PEBAL model is denoted by

$$p_\theta(y|\mathbf{x})_\omega = \frac{\exp(f_\theta(y; \mathbf{x})_\omega)}{\sum_{y' \in \{1, \dots, Y+1\}} \exp(f_\theta(y'; \mathbf{x})_\omega)}, \quad (11.1)$$

where θ is the model parameter, ω indexes a pixel in the image lattice Ω , $p_\theta(y|\mathbf{x})_\omega$ represents the probability of labelling pixel ω with $y \in \{1, \dots, Y + 1\}$, and $f_\theta(y; \mathbf{x})_\omega$ is the logit for class y at pixel ω .

To train the model in (11.1), we formulate a cost function that jointly trains PAL and EBM to classify anomalous pixels. An important training hyper-parameter for PAL is the penalty to abstain from the classification into one of the inlier classes in $\{1, \dots, Y\}$ —this penalty is generally tuned to a single value for all training samples through model selection (e.g., cross validation) [142]. Instead of treating this as a tunable hyper-parameter, we propose the use of EBM (defined below in (11.4)) to automatically estimate this penalty during the training process for each pixel within each training image. More specifically, the cost function to train the PEBAL model in (11.1) is:

$$\begin{aligned} \ell(\mathcal{D}^{in}, \mathcal{D}^{out}, \theta) = & \\ & \sum_{(\mathbf{x}, \mathbf{y}^{in}) \in \mathcal{D}^{in}} (\ell_{pal}(\theta, \mathbf{y}^{in}, \mathbf{x}, E_\theta(\mathbf{x})) + \lambda \ell_{ebm}^{in}(E_\theta(\mathbf{x})) + \ell_{reg}(E_\theta(\mathbf{x}))) + \\ & \sum_{(\mathbf{x}, \mathbf{y}^{out}) \in \mathcal{D}^{out}} (\ell_{pal}(\theta, \mathbf{y}^{out}, \mathbf{x}, E_\theta(\mathbf{x})) + \lambda \ell_{ebm}^{out}(E_\theta(\mathbf{x})) + \ell_{reg}(E_\theta(\mathbf{x}))). \end{aligned} \quad (11.2)$$

where $\ell_{pal}(\cdot)$ denotes the PAL loss defined as

$$\ell_{pal}(\theta, \mathbf{y}, \mathbf{x}, E_\theta(\mathbf{x})) = - \sum_{\omega \in \Omega} \log \left(f_\theta(y_\omega; \mathbf{x})_\omega + \frac{f_\theta(Y + 1; \mathbf{x})_\omega}{a_\omega} \right), \quad (11.3)$$

with $y_\omega \in \{1, \dots, Y\}$ for \mathbf{y}^{in} , $y_\omega \in \{1, \dots, Y + 1\}$ for \mathbf{y}^{out} , and a_ω denotes the pixel-wise penalty associated with abstaining from the classification of the inlier classes. The minimisation of the loss in (11.3) will abstain from classifying outlier pixels into one of the inlier classes, where a pixel is estimated to be an outlier with a_ω . Before formulating a_ω , let us define the inlier free energy at pixel ω , which is denoted by $E_\theta(\mathbf{x})_\omega$ and computed with the *logsumexp* operator as follows [81, 117, 138]:

$$E_\theta(\mathbf{x})_\omega = - \log \sum_{y \in \{1, \dots, Y\}} \exp(f_\theta(y; \mathbf{x})_\omega). \quad (11.4)$$

The pixel-wise penalty associated with abstaining from the classification of the inlier classes is defined by

$$a_\omega = (-E_\theta(\mathbf{x})_\omega)^2, \quad (11.5)$$

which means that the larger the a_ω (i.e., low inlier free energy, so the sample is an inlier), the higher the loss to abstain from classifying into one of the Y classes, and low value of a_ω (i.e., high free inlier energy, which means an outlier sample) implies a lower loss to abstain from classifying one of the Y classes. Also in (11.2), $\ell_{ebm}^{in}(\cdot)$

(weighted by hyper-parameter λ) represents the EBM loss that pushes the inlier free energy in (11.4) for samples in \mathcal{D}^{in} to low values, with

$$\ell_{ebm}^{in}(E_{\theta}(\mathbf{x})) = \sum_{\omega \in \Omega} (\max(0, E_{\theta}(\mathbf{x})_{\omega} - m_{in}))^2, \quad (11.6)$$

representing the loss of having inlier samples with free energy larger than threshold m_{in} , and

$$\ell_{ebm}^{out}(E_{\theta}(\mathbf{x})) = \sum_{\omega \in \Omega} (\max(0, m_{out} - E_{\theta}(\mathbf{x})_{\omega}))^2, \quad (11.7)$$

denoting the loss of having outlier samples with inlier free energy smaller than threshold m_{out} , where the margin losses in (11.6) and (11.7) effectively create an energy gap between normal and abnormal pixels. The last term to define in (11.2) is the inlier free energy regularisation loss to enforce that anomalous pixels are sparse and pixel anomaly classification is smooth (i.e., anomalous pixels tend to have anomalous neighbouring pixels), which is defined as

$$\ell_{reg}(E_{\theta}(\mathbf{x})) = \sum_{\omega \in \Omega} \beta_1 |E_{\theta}(\mathbf{x})_{\omega} - E_{\theta}(\mathbf{x})_{\mathcal{N}(\omega)}| + \beta_2 |E_{\theta}(\mathbf{x})_{\omega}|, \quad (11.8)$$

where β_1 and β_2 are hyper-parameters that weight the contributions of the smoothness and sparsity and sparsity regularisations, and $\mathcal{N}(\omega)$ denotes neighbouring pixels in horizontal and vertical directions.

11.3.3 Training and Inference

Training. An important point of the training process is how to setup the inlier and outlier datasets \mathcal{D}^{in} and \mathcal{D}^{out} . A recently published paper [27] carefully selects images to be included in \mathcal{D}^{out} by making sure that the segmentation labels presented in those images do not overlaps with the inlier labels. In particular for [27], \mathcal{D}^{in} has images and annotations from Cityscape and \mathcal{D}^{out} has images and annotations from COCO [129]. We argue that there are two issues with this strategy to form \mathcal{D}^{out} , which are: 1) the selected COCO images generally only contain anomalous pixel labels, leading to unstable training of the outlier losses (i.e., second summation in (11.2)) given the exclusive presence of the anomaly class (in effect, this becomes a one-class segmentation problem); 2) re-training the model with images containing only anomalous pixels removes the semantic context of inlier pixels when training for the outlier losses, which can deteriorate the segmentation accuracy of the inlier labels.

To mitigate these issues, we form \mathcal{D}^{out} using a novel extension based on CutMix and CutPaste [120, 256], which we refer to as AnomalyMix. AnomalyMix cuts the anomalous objects from an outlier dataset (e.g., COCO) using its labelled masks and paste them into the images of the inlier dataset (e.g., CitySpace), where we label the

pixels of the anomalous object with the class $Y + 1$ – these images are then inserted into \mathcal{D}^{out} . AnomalyMix addresses the two issues above because the outlier images now contain a combination of inlier and outlier pixels, allowing a balanced learning and keeping the visual context of inlier labels when training for the outlier losses. Furthermore, AnomalyMix can form a potentially infinite number of training images for \mathcal{D}^{out} given the range of transformations to be applied to the cut objects and the locations of the inlier images that the objects can be pasted. Previous papers [49, 103] argue that re-training the whole segmentation model can jeopardise the segmentation accuracy for the inlier classes. Furthermore, such re-training requires a long training time, leading to inefficient optimisation. In this work, we propose to **fine-tune only the final classification block** using the loss in (11.2), instead of re-training the whole segmentation model. Besides being efficient, this fast fine-tuning keeps the segmentation accuracy of the model in the original dataset used for pre-training the model. Furthermore, an interesting side-effect of our training is that the cost function in (11.2) will calibrate the segmentation prediction for the inlier classes. This happens because the terms $\ell_{pal}(\cdot)$, $\ell_{ebm}^{in}(\cdot)$ and $\ell_{ebm}^{out}(\cdot)$ jointly constrain the maximisation of logits and naturally calibrate classification confidence.

Inference. During inference, pixel-wise anomaly detection is performed by computing the inlier free energy score $\mathbb{E}_\theta(\mathbf{x})_\omega$ from (11.4) for each pixel position ω given a test image \mathbf{x} and inlier segmentation is obtained from the inlier classes from the PEBAL model in (11.1). Following [103], we also apply a Gaussian smoothing kernel to produce the final energy map.

11.4 Experiment

11.4.1 Datasets

LostAndFound [179] is one of the first publicly available urban driving scene anomaly detection datasets containing real-world anomalous objects. The dataset has an official testing set containing 1,203 images with small obstacles in front of the cars, collecting from 13 different street scenes, featuring 37 different types of anomalous objects with various sizes and material.

Fishyscapes [20] is a high-resolution dataset for anomaly estimation in semantic segmentation for urban driving scenes. The benchmark has an online testing set that is entirely unknown to the methods. The dataset is composed by two data sources: Fishyscapes LostAndFound that contains a set of real road anomalous objects [179] and a blending-based Fishyscapes Static dataset. The Fishyscapes LostAndFound validation set consists of 100 images from the aforementioned LostAndFound dataset with refined labels and the Fishyscapes Static validation set contains 30 images with the blended anomalous objects from Pascal VOC [69]. For all datasets, we select the checkpoints based on the results on the public validation sets, but submitted our code

and checkpoints to the benchmark website to be evaluated on their hidden test sets.

Road Anomaly [130] contains real-world road anomalies in front of the vehicles. The dataset has 60 images from the Internet, containing unexpected animals rocks, cones and obstacles. Unlike the LostAndFound and Fishyscapes, this dataset contains abnormal objects with various scales and sizes, making it even more challenging.

11.4.2 Implementation Details

Following [26, 27], we use DeepLabv3+ [28] with WideResNet34 trained by Nvidia [271] and ResNet101 from [103] as the backbone of our segmentation models. The training details of those models can be found in their original papers or our supplementary material. The models are trained on Cityscapes [39] training set. For our PEBAL fine-tuning, we empirically set the m_{in} and m_{out} in Eq. 11.6 and Eq. 11.7 as -12 and -6, respectively. The weights β_1 and β_2 in Eq. 11.8 are set to $5e - 4$ and $3e - 6$ [211], λ in Eq. 11.2 to 0.1, and the weight of ℓ_{ebm} to 0.1, respectively. Note that those hyper-parameters are selected at the first training epoch to normalise loss values to a similar scale. We also show our model can obtain consistently SOTA results regardless of the selection of hyper-parameters in the supplementary material. Our training consists of fine-tuning the final classification block of the model for 20 epochs. We use the same resolution of random crop as in [271], and use Adam with a learning rate of $1e^{-5}$. The batch size is set to 16. Following [27], for our AnomalyMix augmentation, we randomly sample 297 images as training data from the remaining COCO images that do not contain objects in Cityscapes or our anomaly validation/testing sets and randomly apply AnomalyMix to mix them into the Cityscape training images, following Chan et al. [27].

11.4.3 Evaluation Measures

Following [20, 27, 49, 103], we compute the the area under receiver operating characteristics (AUROC), average precision (AP), and the false positive rate at a true positive rate of 95% (FPR95) to validate our approach. For Fishyscapes public leaderboard, we use AP and FPR95 to compare with other methods, same as their website.

11.4.4 Comparison on Anomaly Segmentation Benchmarks

Comparison on LostAndFound.

Table 11.1 shows the result on the testing set of LostAndFound. Notably, our approach surpasses the previous baseline approaches (i.e., MSP [93], Mahalanobis [119], Max Logit [95] and Entropy [95]) by 10% to 40% AP, and 13% to 22% FPR95, respectively. When compared with previous SOTA approaches such as SynBoost [49], SML [103] and Meta-OoD [27], we improve the AP performance by a large margin (15% to 40%), and decrease the FPR95 by about 5% to 70%. This illustrates the robustness and effectiveness on detecting small and distant anomalous objects given that the

Table 11.1: Anomaly segmentation results on **LostAndFound** testing set, with **WideResnet34** backbone. All methods use the same segmentation models. * indicate that the model requires additional learnable parameters. † indicates that the results are obtained from the official code with our WideResnet34 backbone.

Methods	AUC \uparrow	AP \uparrow	FPR ₉₅ \downarrow
MSP [93]	85.49	38.20	18.56
Mahalanobis [119]	79.53	42.56	24.51
Max Logit [95]	94.52	65.45	15.56
Entropy [95]	86.52	50.66	16.95
Energy [138]	94.45	66.37	15.69
Meta-OoD [27]	97.95	71.23	5.95
†SML [103]	88.05	25.89	44.48
†SynBoost* [49]	98.38	70.43	4.89
Deep Gambler [142]	98.67	72.73	3.81
Ours	99.76	78.29	0.81

dataset contains mostly real-world small objects. Our PEBAL also improves the EBM baseline [138] and the AL baseline based on Deep Gambler [142]. This demonstrates that a simple adaptation of AL and EBM is not enough to enable accurate pixel-wise anomaly detection. Previous SOTA SML [103] aims to balance the inlier class-wise discrepancy on prediction scores, which is disadvantageous for measuring performance on LostAndFound test set since there may be no classes in the evaluation other than the road class (i.e., most of the inlier classes within LF test set is road class), thus leading to significant performance variations between LostAndFound and Fishyscapes. It is worth noting that our approach achieves 1.03% FPR95, significantly reducing the false positive pixels, improving the chances of applying it to real-world applications.

Comparison on Fishyscapes Leaderboard.

Table 11.2 shows the leaderboard results on the test set of Fishyscapes LostAndFound and Fishyscapes Static. Following [103], we compared the methods based on whether they require re-training of the entire segmentation network, adding the extra network, or utilising the OoD data. We achieve the SOTA performance by a large margin on Fishyscapes leaderboard when compared with the previous methods except [17] (Static) that rely on an inefficient re-training segmentation model, extra learnable parameters, and extra OoD training data. Without re-training the entire network or adding extra learnable parameters, our approach can work efficiently to surpass previous SOTA competing approaches that fall into the same category by about 13% to 42% on LostAndFound and 40% to 50% AP on Static. Such significant improvements indicate the generalisation ability of our proposed PEPAL on detecting a wide variety of unseen abnormalities (i.e., of different size, type, scene, and distance) substantially reducing false negative and positive pixels. Moreover, it is worth noting that PEBAL

Table 11.2: Comparison with previous approaches on **Fishyscapes Leaderboard**. We achieve a new state-of-the-art performance among the approaches that require extra OoD data, and without re-training the segmentation networks and extra networks on Fishyscapes Leaderboard.

Models	re-training	Extra Network	OoD Data	FS LostAndFound		FS Static	
				AP \uparrow	FPR95 \downarrow	AP \uparrow	FPR95 \downarrow
Discriminative Outlier Detection Head [17]	\checkmark	\checkmark	\checkmark	31.31	19.02	96.76	0.29
MSP [93]	\times	\times	\times	1.77	44.85	12.88	39.83
Entropy [95]	\times	\times	\times	2.93	44.83	15.41	39.75
SML [103]	\times	\times	\times	31.05	21.52	53.11	19.64
kNN Embedding - density [20]	\times	\times	\times	3.55	30.02	44.03	20.25
Bayesian Deeplab [155]	\checkmark	\times	\times	9.81	38.46	48.70	15.05
Density - Single-layer NLL [20]	\times	\checkmark	\times	3.01	32.9	40.86	21.29
Density - Minimum NLL [20]	\times	\checkmark	\times	4.25	47.15	62.14	17.43
Image Resynthesis [130]	\times	\checkmark	\times	5.70	48.05	29.6	27.13
OoD Training - Void Class	\checkmark	\times	\checkmark	10.29	22.11	45.00	19.40
Dirichlet Deeplab [149]	\checkmark	\times	\checkmark	34.28	47.43	31.30	84.60
Density - Logistic Regression [20]	\times	\checkmark	\checkmark	4.65	24.36	57.16	13.39
SynBoost [49]	\times	\checkmark	\checkmark	43.22	15.79	72.59	18.75
Ours	\times	\times	\checkmark	44.17	7.58	92.38	1.73

Table 11.3: Anomaly segmentation results on **Fishyscapes validation sets** (LostAndFound and Static), and the **Road Anomaly testing set**, with **WideResnet34** backbone. * indicate that the model requires additional learnable parameters. \dagger indicates that the results are obtained from the official code with our WideResnet34 backbone. Best and second best results in bold.

Methods	FS LostAndFound			FS Static			Road Anomaly		
	AUC \uparrow	AP \uparrow	FPR ₉₅ \downarrow	AUC \uparrow	AP \uparrow	FPR ₉₅ \downarrow	AUC \uparrow	AP \uparrow	FPR ₉₅ \downarrow
MSP [93]	89.29	4.59	40.59	92.36	19.09	23.99	67.53	15.72	71.38
Max Logit [93]	93.41	14.59	42.21	95.66	38.64	18.26	72.78	18.98	70.48
Entropy [95]	90.82	10.36	40.34	93.14	26.77	23.31	68.80	16.97	71.10
Energy [138]	93.72	16.05	41.78	95.90	41.68	17.78	73.35	19.54	70.17
Mahalanobis [119]	96.75	56.57	11.24	96.76	27.37	11.7	62.85	14.37	81.09
Meta-OoD [49]	93.06	41.31	37.69	97.56	72.91	13.57	-	-	-
\dagger Synboost* [49]	96.21	60.58	31.02	95.87	66.44	25.59	81.91	38.21	64.75
\dagger SML [103]	94.97	22.74	33.49	97.25	66.72	12.14	75.16	17.52	70.70
Deep Gambler [142]	97.82	31.34	10.16	98.88	84.57	3.39	78.29	23.26	65.12
Ours	98.96	58.81	4.76	99.61	92.08	1.52	87.63	45.10	44.58

reduces the amount of false positive pixels to 7.58 and 1.73 FPR on the two datasets. This result is publicly available on the Fishyscapes website.

Comparison on Fishyscapes validation sets and Road Anomaly.

In Tables 11.3 and 11.4, we compare our approach on the Fishyscapes validation sets and Road Anomaly using two different backbones. Our model outperforms the previous methods by a large margin on all three benchmarks, regardless of the backbones and their segmentation accuracy. To verify the applicability of our method, except for the modern WideResnet34 backbone, we use a ResNet101 DeepLabv3+ to investigate the performance in terms of the size of the architecture and its inlier segmentation

Table 11.4: Anomaly segmentation results on **Fishyscapes validation sets** (LostAndFound and Static), and the **Road Anomaly testing set**, with **Resnet101** backbone. * indicate that the model requires additional learnable parameters. † indicates that the results are obtained from the official code with our Resnet101 backbone. Best and second best results in bold.

Methods	FS LostAndFound			FS Static			Road Anomaly		
	AUC ↑	AP ↑	FPR ₉₅ ↓	AUC ↑	AP ↑	FPR ₉₅ ↓	AUC ↑	AP ↑	FPR ₉₅ ↓
MSP [93]	86.99	6.02	45.63	88.94	14.24	34.10	73.76	20.59	68.44
Max Logit [93]	92.00	18.77	38.13	92.80	27.99	28.50	77.97	24.44	64.85
Entropy [95]	88.32	13.91	44.85	89.99	21.78	33.74	75.12	22.38	68.15
Energy [138]	93.50	25.79	32.26	91.28	31.66	37.32	78.13	24.44	63.36
†SynthCP* [246]	88.34	6.54	45.95	89.9	23.22	34.02	76.08	24.86	64.69
†Synboost* [49]	94.89	40.99	34.47	92.03	48.44	47.71	85.23	41.83	59.72
SML [103]	96.88	36.55	14.53	96.69	48.67	16.75	81.96	25.82	49.74
Deep Gambler [142]	97.19	39.77	12.41	97.51	67.69	15.39	85.45	31.45	48.79
Ours	99.09	59.83	6.49	99.23	82.73	6.81	92.51	62.37	28.29

accuracy. The results demonstrate that our approach is applicable to a wide-range of segmentation models, indicating the effectiveness of PEBAL to adapt to real-world systems.

Moreover, our fine-tuning sacrifices only marginally the inlier segmentation accuracy (i.e., 0.2% - 0.7% mIoU on Cityscapes) for both backbones, achieving good performance on both inlier and anomaly segmentation. We present details of all inlier segmentation models (i.e., Cityscapes training setup and mIoU), and include more experimental results of other DeepLabv3+ checkpoints in supplementary material.

Remarks – Superior Performance on Challenging Benchmarks.

Each dataset has different challenges. For example, the LostAndFound testing set considers only drivable areas with homogeneous normal scenes (i.e., road) and limited categories of abnormalities (i.e., road obstacles), leading to a relatively less challenging benchmark on which most methods can obtain good AUC performance, as shown in Tables 11.1, 11.3 and 11.4. On the contrary, Fishyscapes and RoadAnomaly contain large number of heterogeneous inlier and outlier pixels from diverse classes, leading to significantly more difficult testbeds than the LostAndFound testing set. Furthermore, Fishyscapes and RoadAnomaly contain domain shift compared with Cityscapes (e.g., both datasets contain different scenes than Cityscapes) and have different types/sizes of OoD objects. Most existing SOTA methods work ineffectively on these two datasets due to those challenges, while our adaptive pixel-level anomaly class learning helps our model effectively detect these challenging inlier and outlier pixels in the aforementioned heterogeneous and domain-shifted scenes, yielding substantial improvements (i.e., 20% to 50%) to previous approaches, as shown in Tables 11.2, 11.3 and 11.4.

Table 11.5: Ablation studies for anomaly segmentation on **LostAndFound**, with **WideResnet34** backbone, where all proposed modules are trained with COCO OE images with AnomalyMix. CE denotes the baseline method that adds an extra OoD class to learn the OE training samples with cross-entropy (first row).

CE	ℓ_{ebm}	ℓ_{pal}	ℓ_{reg}	AUC \uparrow	AP \uparrow	FPR ₉₅ \downarrow
✓				96.88	69.02	8.03
	✓			97.88	70.24	8.92
		✓		98.67	72.73	3.81
	✓	✓		99.63	77.19	1.19
	✓	✓	✓	99.76	78.29	0.81

Table 11.6: The performance comparison of our approach on Fishyscapes benchmark w.r.t different **diversity of OE classes** (mean results over six random seeds), in terms of AP and FPR95.

Class Per.	FS LostAndFound		FS Static	
	AP \uparrow	FPR ₉₅ \downarrow	AP \uparrow	FPR ₉₅ \downarrow
1%	53.57 \pm 3.74	6.97 \pm 1.98	85.84 \pm 1.01	3.05 \pm 0.97
5%	52.16 \pm 3.88	6.58 \pm 1.95	90.57 \pm 1.75	1.93 \pm 0.52
10%	55.14 \pm 3.02	5.78 \pm 1.59	91.37 \pm 1.28	1.64 \pm 0.58
25%	55.48 \pm 3.32	5.98 \pm 1.27	91.28 \pm 1.94	1.77 \pm 0.18
50%	56.69 \pm 2.57	5.32 \pm 1.16	91.88 \pm 0.71	1.62 \pm 0.05
75%	57.86 \pm 2.83	5.11 \pm 1.69	91.85 \pm 0.56	1.63 \pm 0.09

11.4.5 Ablation Study

Table 11.5 shows the contribution of each component of our PEBAL on the LostAndFound testing set. All modules are trained with COCO OE images using AnomalyMix. Adding an extra OoD class to learn the OE training samples with cross-entropy (**CE**) is our baseline (first row). To justify the effectiveness of our proposed joint training, we show the results using energy-based models (ℓ_{ebm} without ℓ_{pal}) and pixel-wise abstention (ℓ_{pal} with pre-defined fixed penalty). Both outperform the CE baselines (AP=70.2, FPR=8.9 and AP=72.7, FPR=3.8 vs. AP=69, FPR=8.03), while our proposed joint training ($\ell_{ebm} + \ell_{pal}$) obtains 77.19% of AP and 1.19% of FPR, improving over each module by 4% to 7%. This indicates the effectiveness of our joint training and the significance of our proposed PAL with learnable adaptive energy-based penalties a_ω . Finally, the smoothness and sparsity regularisation losses stabilise the training and further improve the performance.

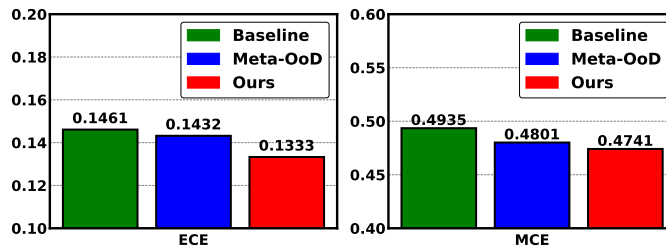


Figure 11.3: Confidence calibration performances between WideResnet34 baseline, Meta-OoD [27], and our approach.

11.4.6 Outlier Samples, Calibration and Efficiency

Outlier Diversity and Efficiency.

In Table 11.6, we randomly select 1%, 5%, 10%, 25% 50%, and 75% of COCO classes as the OE data during training and compute the mean results over six different random seeds. We achieve consistent AP and FPR performance regardless of the number of COCO classes used during the training on Fishyscapes. It is also worth noting that our approach can effectively learn the PEBAL model using **only one class** (1% in Table 11.6) of outlier data, which selects some of the irrelevant classes of COCO objects that are not possible to be found on road in real life (e.g., dining table, laptop, and clock). The results indicate that our model can consistently achieve SOTA performance on Fishyscapes without a careful selection of OE classes, demonstrating the robustness of our approach under diverse outlier classes. We also investigate the outlier sample efficiency of our model w.r.t smaller OE training sets with a fixed 100% COCO classes (80 classes) on Fishyscapes in Table 11.7, and we achieve consistently good performance regardless the number of outlier training samples. All those experiments show the applicability of our PEBAL to real-world autonomous driving systems.

Confidence Calibration.

In Fig. 11.3, we show that our model can also improve the calibration of the segmentation confidence. This figure shows that we improve the ECE and MCE [84] scores by a small margin, showing another benefit of using our PEBAL approach.

Computational Efficiency.

We compare the computational efficiency of our PEBAL with previous SOTA Meta-OoD [27] and Synboost [49] in terms of the trainable parameters, training time and mean inference time per image, on an NVIDIA3090. As PEBAL requires the fine-tuning of the final classification block, it has only 1.3M parameters and each training epoch takes about 12 minutes, which is significantly less than the re-training approach Meta-OoD that has 137.1M parameters and each training epoch takes about 26 minutes, and the reconstruction based approach Synboost that takes about 33 minutes to train a epoch of its re-synthesis and dissimilarity networks with 157.3M parameters. Moreover,

Table 11.7: The performance comparison of our approach on Fishyscapes benchmark w.r.t different **amount of OE training samples** (mean results over six random seeds), in terms of AP and FPR95.

Train Size	FS LostAndFound		FS Static	
	AP \uparrow	FPR ₉₅ \downarrow	AP \uparrow	FPR ₉₅ \downarrow
5%	54.32 \pm 1.89	5.77 \pm 2.38	89.11 \pm 1.52	2.23 \pm 0.65
10%	56.28 \pm 1.05	4.66 \pm 1.36	90.02 \pm 0.57	1.67 \pm 0.28
25%	56.18 \pm 1.69	4.81 \pm 1.44	91.23 \pm 0.95	1.63 \pm 0.22
50%	57.34 \pm 1.19	4.75 \pm 1.32	91.29 \pm 0.92	1.67 \pm 0.17

our method also has a much faster mean inference time of 0.55s compared to 0.68s of Meta-OoD and 1.95s of Synboost. Those results suggest the practicability of our model in real-world self-driving systems.

11.5 Conclusions and Discussions

We proposed a simple yet effective approach, named Pixel-wise Energy-biased Abstention Learning (PEBAL), to fine-tune the last block of a segmentation model to detect unexpected road anomalies. The approach introduces a non-trivial training that jointly optimises a novel pixel-wise abstention learning and an energy-based model to learn an adaptive pixel-wise anomaly class, in which a new pixel-wise energy-biased penalty estimation method is proposed to improve the precision and robustness to detect small and distant anomalous objects. The resulting model significantly reduces the false positive and false negative detected anomalies, compared with previous SOTA methods. The results on four benchmarks demonstrate the accuracy and robustness of our approach to detect anomalous objects regardless of the amount or diversity of exposed training outliers. Despite the remarkable performance on most datasets, PEBAL is not as effective on the most challenging dataset, Road Anomaly, that contains significantly more diverse and realistic anomalous objects. We plan to further enhance the generalisation of our model to accurately detect more unknown, diverse anomalies.

Chapter 12

Conclusion

In this thesis, we developed effective deep anomaly detection methods for computer vision and medical image analysis tasks. First, we discussed the issues of current reconstruction-based UAD approaches. For instance, UAD models often generalise so well that they can also accurately reconstruct abnormalities (i.e., with low reconstruction error), leading to potential mis-detection of anomalies. To address this issue, we proposed ADGAN for anomaly detection using a dual GAN structure that provides stronger constraints to map between the input image and GAN’s latent spaces.

Inspired by the recently proposed Masked Autoencoders [89] (MAE), we introduced a new UAD reconstruction method to address the aforementioned low-reconstruction error issue, named MemMC-MAE, for anomaly detection and localisation in medical images. To the best of our knowledge, our MemMC-MAE is the first UAD method based on MAE. We modified the standard MAE transformer encoder with a novel memory-augmented self-attention operator and a new multi-level cross-attention for the MAE transformer decoder. MemMC-MAE randomly masked large regions of the input image during its reconstruction, reducing the risk that it will produce low reconstruction errors because anomalous regions are likely to be masked and cannot be reconstructed. However, when anomaly regions are not masked, then the normal patterns stored in the encoder’s memory combined with the decoder’s multi-level cross-attention will constrain the reconstruction ability of the model and enforce it to reconstruct the abnormal images into their normal version. The resulting model showed SOTA anomaly detection and localisation performance on colonoscopy and Covid-19 Chest X-ray datasets.

We then proposed the first few-shot anomaly detection framework, named FSAD-NET. This new approach introduces a middle ground between the imbalanced learning methods that generally require a relatively large amount of abnormal training data and the UAD methods that use no abnormal training data. FSAD-NET is trained to learn fine-grained anomalous features from only a few abnormal samples and to generalise well for unseen anomalies. The resulting FSAD-NET achieves significantly

better accuracy than previous UAD and imbalanced learning methods.

Another major challenge faced by UAD methods is how to learn effective low-dimensional feature representations to detect and localise subtle abnormal lesions (e.g., small and flat polyp). Such low-dimensional representations are crucial for downstream anomaly classifiers. Despite the recent progress of self-supervised methods [10, 30, 77, 90, 94, 133] that have been shown to learn effective representations for general computer vision tasks [10, 77, 94], the effectiveness of those methods in UAD for medical images remains unexplored. We addressed this issue with a new self-supervised pre-training solution for UAD methods, named constrained contrastive distribution (CCD) learning, which enforces non-uniform representation distribution by constraining contrastive distribution learning with two pretext tasks. Our CCD achieves state-of-the-art results when pre-training a wide range of off-the-shelf anomaly classifiers, indicating that our method is agnostic to downstream classifiers. Although achieving good results in many benchmarks, the contrastive learning explored by CCD ignores the fact that the downstream UAD methods need to recognise one class of normal images and a small number of sub-classes of disease images. Moreover, CCD’s data augmentation is based on methods designed for computer vision images and cannot produce realistic synthesised medical image anomalies. Those two issues can challenge the training of downstream UAD approaches for medical image analysis problems. Hence, we extended our CCD pre-training approach by introducing a new contrastive learning loss to convert the training of one-class classifiers into the training of multi-class clustering methods, with the goal of constructing denser and tighter clusters with the proposed MedMix data augmentation to simulate realistic medical abnormalities. The resulting MSACL pre-training yields significantly better performance than our previous CCD pre-training. To the best of our knowledge, Our CCD and MSACL are the first works to explore self-supervised pre-training for UAD tasks in medical image analysis.

Despite the remarkable performance of UAD methods with self-supervised pre-training, the training of anomaly classifier often suffers from overfitting the training data, especially when the training set is small or contaminated with anomalous samples. To address this issue, we proposed a novel UAD model, named interpolated Gaussian descriptor (IGD), to perform unsupervised anomaly detection and segmentation. Our IGD aims to tackle the overfitting and unstable training issues (e.g., catastrophic collapse) of previous OCC/UAD models by formulating the optimisation as an EM algorithm. We show that the proposed one-class Gaussian anomaly classifier trained with adversarially interpolated samples enables a robust representation of normal samples, which is able to achieve the best performance on six anomaly detection datasets. To the best of our knowledge, IGD is also the first method to assess model’s robustness under insufficient and contaminated datasets.

For weakly supervised video anomaly detection, we introduced a novel method, named Robust Temporal Feature Magnitude learning (RTFM). RTFM learns a tem-

poral feature magnitude mapping function that: 1) detects rare abnormal snippets from abnormal videos containing many normal snippets, and 2) guarantees a large margin between normal and abnormal snippets. The proposed RTFM model is able to learn discriminative features that allows a robust identification of subtle anomalies from videos labelled as abnormal, allowing a better exploitation of abnormal training data. We then adapted this WVAD setup to polyp frame detection with a novel convolutional transformer-based multiple instance learning method designed to identify abnormal polyps frames from anomalous videos containing at least one frame with polyp. Our transformer architecture can seamlessly map the local and global temporal dependencies while simultaneously optimising video and snippet-level anomaly scores. Moreover, a novel and effective contrastive snippet mining (CSM) was proposed to enable the selection of challenging abnormal polyp frames from abnormal colonoscopy videos. Our model showed substantially better results than previous SOTA competing methods on our newly proposed large-scale colonoscopy video dataset.

Finally, we introduced a simple yet effective approach, named Pixel-wise Energy-biased Abstention Learning (PEBAL), to tackle anomaly segmentation tasks. This task is particularly important for self-driving systems because current segmentation models often fail to detect unexpected road anomalies, leading to potentially fatal traffic collisions. We argue that previous uncertainty-based approaches tend to not work well for pixel-wise tasks, and reconstruction-based approaches depend on an extra reconstruction network that is hard to train and inefficient to run in real-time self-driving embedded systems. Hence, to resolve the issues above, our PEBAL introduced a non-trivial training that learns a novel adaptive energy-based penalty for every pixel through an energy-based model, which is jointly optimised with a novel pixel-wise abstention learning. The resulting method is efficient to run and able to significantly improve the precision and robustness of the detection of small and distant anomalous objects, compared with previous competing approaches on four benchmarks.

12.1 Limitations and Future Work

In this thesis, we developed new methods for anomaly detection under the unsupervised, weakly supervised and few-shot settings. However, in real-world applications, the normal datasets may contain a few anomalous samples, which highlights the importance of studying methods that can be trained with anomaly contaminated training data. Hence, we plan to develop new approaches that will be robust to around 5% to 10% of anomalous data incorrectly present in the training set of normal data. Our methods will be designed to identify such anomalous data by pseudo-labelling them during the training process. Unlike our previous IGD approach that showed robustness under anomaly contamination, our planned approach will take advantage of the outlier samples through the pseudo labelling process and re-formulate the problem as

an open-set noisy-label learning.

For self-supervised pre-training of UAD methods in medical images, the current CCD and MSACL models only consider 2D images. In future work, we plan to extend such self-supervised pre-training approaches to 3D CT and MRI images. Moreover, inspired by the recent success of vision transformer, we will further explore the self-supervised pre-training for UAD methods using the transformer architecture.

IGD proposed a Gaussian anomaly classifier to learn normality patterns from whole images only, ignoring the important pixel/patch level information. In future work, we plan to study the use of Gaussian anomaly classifier using pixel/patch data to enable better segmentation accuracy. Furthermore, inspired by the MAE pre-training approach for multi-class classification [89], we will adapt our MemMC-MAE pre-training to other UAD methods in future work.

Despite the remarkable performance on most datasets, PEBAL is not as effective on the most challenging dataset, Road Anomaly, that contains significantly more diverse and realistic anomalous objects under domain-shifted scenes. In the future, we will improve PEBAL’s robustness to domain shift scenarios with data transformations (e.g., color jittering and Gaussian blur) to suppress domain-shifted feature correlations, and a combination of instance and batch normalisation.

For weakly-supervised video anomaly detection (WVAD) approaches, the two models introduced in this thesis (in Chapter 8 and 9) will fail on online applications because the models require the whole videos to be analysed during the self-attention operation. To resolve this issue, we plan to propose an online approximator to compute the self-attention using only past video snippets. Another potential future work for the WVAD setup is on the exploration of unsupervised deep clustering approaches to assign video level pseudo labels during training, completely avoiding the laborious annotation process. Lastly, low computational complexity is critical to enable the deployment of systems to real-time applications, such as video surveillance and self-driving cars. Therefore, we plan to modify our anomaly detection systems so that they have small run-time and memory complexities. We hope this thesis can serve as inspiration to build effective, robust, and practical anomaly detectors for the computer vision community.

Appendix A

IGD (Chapter 6) Appendix

A.1 Datasets

CIFAR10 contains 60,000 images with 10 classes. MNIST and Fashion MNIST contain 70,000 images with 10 classes of handwritten digits and fashion products, respectively. MVTec AD [13] contains 5,354 high-resolution real-world images of 15 different industry object and textures. The normal class of MVTec AD is formed by 3,629 training and 467 testing images without defects. The anomalous class has more than 70 categories of defects (such as dents, structural fails, contamination, etc.) and contains 1,258 testing images. MVTec AD provides pixel-wise ground truth annotations for all anomalies in the testing images, allowing the evaluation of anomaly detection and localisation. Hyper-Kvasir has 1,600 normal images without polyps in the training set and 500 in the testing set; and 1,000 abnormal images containing polyps in the testing set. For LAG, we have 2,343 normal images without glaucoma in the training set; and 800 normal images and 1,711 abnormal images with glaucoma for testing.

A.2 Global and Local IGD Models

Figure A.1 shows an example of a multi-scale structural and non-structural anomaly localisation result for an MVTec AD image, using both the local and global IGD models.

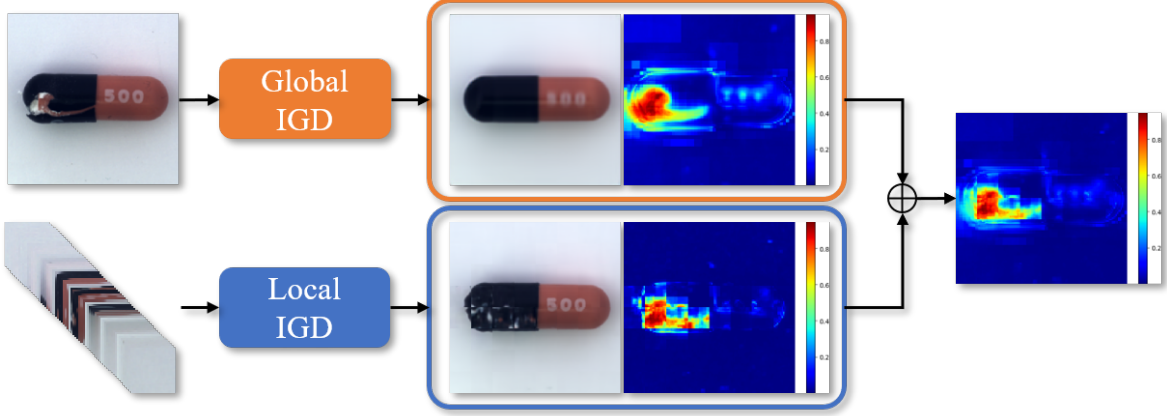


Figure A.1: Example of the multi-scale structural and non-structural anomaly localisation result for an MVTEC AD [13] image, using both the local and global IGD models. The global model tends to produce smooth results but with some mistakes, while the local model produces jagged results, but without the global mistakes, so by combining the two results, we obtain a smooth and correct anomaly heatmap.

A.3 Multi-scale Structure Similarity Index (MS-SSIM) Score

The MS-SSIM loss uses the MS-SSIM global score, defined as

$$m^{(G)}(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega)) = [l_M(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega))]^{\alpha_M} \times \prod_{m=1}^{m^{(G)}} [c_m(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega))]^{\beta_m} [s_m(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega))]^{\gamma_m}, \quad (\text{A.1})$$

where $\mathbf{x}(\omega)$ denotes an image patch centred at $\omega \in \Omega$ of size $11 \times 11 \times 3$,

$$l_M(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega)) = \frac{2\mu_{\mathbf{x}(\omega)}\mu_{\hat{\mathbf{x}}(\omega)} + C_1}{\mu_{\mathbf{x}(\omega)}^2 + \mu_{\hat{\mathbf{x}}(\omega)}^2 + C_1}, \quad (\text{A.2})$$

$$c_m(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega)) = \frac{2\sigma_{\mathbf{x}(\omega)}\sigma_{\hat{\mathbf{x}}(\omega)} + C_2}{\sigma_{\mathbf{x}(\omega)}^2 + \sigma_{\hat{\mathbf{x}}(\omega)}^2 + C_2}, \quad (\text{A.3})$$

$$s_m(\mathbf{x}(\omega), \hat{\mathbf{x}}(\omega)) = \frac{\sigma_{\mathbf{x}(\omega)\hat{\mathbf{x}}(\omega)} + C_3}{\sigma_{\mathbf{x}(\omega)}\sigma_{\hat{\mathbf{x}}(\omega)} + C_3}, \quad (\text{A.4})$$

with C_1, C_2, C_3 representing pre-defined constants, $\mu_{\mathbf{x}(\omega)}$ denoting the mean intensities of $\mathbf{x}(\omega)$, $\sigma_{\mathbf{x}(\omega)}^2$ the variance of $\mathbf{x}(\omega)$, and $\sigma_{\mathbf{x}(\omega)\hat{\mathbf{x}}(\omega)}$ the covariance of $\mathbf{x}(\omega)$ and $\hat{\mathbf{x}}(\omega)$. In (A.1), $m^{(G)} = 5$ denotes the number of scales, $\beta_1 = \gamma_1 = 0.0448$, $\beta_2 = \gamma_2 = 0.2856$,

$\beta_3 = \gamma_3 = 0.3001$, $\beta_4 = \gamma_4 = 0.2363$, $\alpha_5 = \beta_5 = \gamma_5 = 0.1333$ [242]. We follow $C_i = (K_i L)^2$ (for $i \in \{1, 2, 3\}$) according to [240] and define $L = 4.7579$ as the pixel range with $K_1 = 0.01$, $K_2 = 0.03$ and $C_3 = C_2/2$.

The local score $m^{(L)}(\mathbf{x}^{(L)}(\omega), \hat{\mathbf{x}}^{(L)}(\omega))$ is defined in the same way as in (A.1), where $\mathbf{x}^{(L)}(\omega)$ is an image patch centred at $\omega \in \Omega$ of size $3 \times 3 \times 3$, $m^{(L)} = 4$ scales with weights $\beta_1 = \gamma_1 = 0.0516$, $\beta_2 = \gamma_2 = 0.3295$, $\beta_3 = \gamma_3 = 0.3463$, $\alpha_4 = \beta_4 = \gamma_4 = 0.2726$ modified based on the original proportion for $m^{(G)} = 5$.

A.4 Implementation Details

For this SSL pre-training, we use the SGD optimiser with a learning rate of 0.01, weight decay 10^{-1} , batch size of 32, and 2,000 epochs. Once we obtain the pre-trained encoder with SSL, we remove the MLP layer and attach a linear layer to the backbone with fixed parameters. Note that this SSL is trained from scratch. In contrast to the vanilla self-supervised learning [29] suggesting large batch size, we notice that a medium batch size yields significantly better performance for unsupervised anomaly detection.

For the ImageNet KD pre-training, we minimise the ℓ_2 norm between the 512-dimensional feature vector output from encoder and an intermediate layer of the ImageNet pre-trained ResNet18 [92] with the same 512-dimensional features. For this ImageNet KD pre-training, we use the Adam optimiser with a learning rate of 0.0001, weight decay 10^{-5} , batch size of 64, and 50,000 iterations. Once we obtain the pre-trained encoder of KD, we fix the network parameters and attach a linear layer to reduce the dimensionality of the feature space to 128.

A.5 Visualisation of the Distribution of Testing Samples

Figure A.2 shows the distribution of testing samples in the representation space, using the t-SNE visualisation, for DSVDD [191], Gaussian anomaly classifier (GAC), and our IGD. Notice that the normal samples seem to be more compactly represented with fewer anomalous samples appearing inside the normal cluster. This suggests that IGD has a superior normality description, compared with DSVDD and GAC.

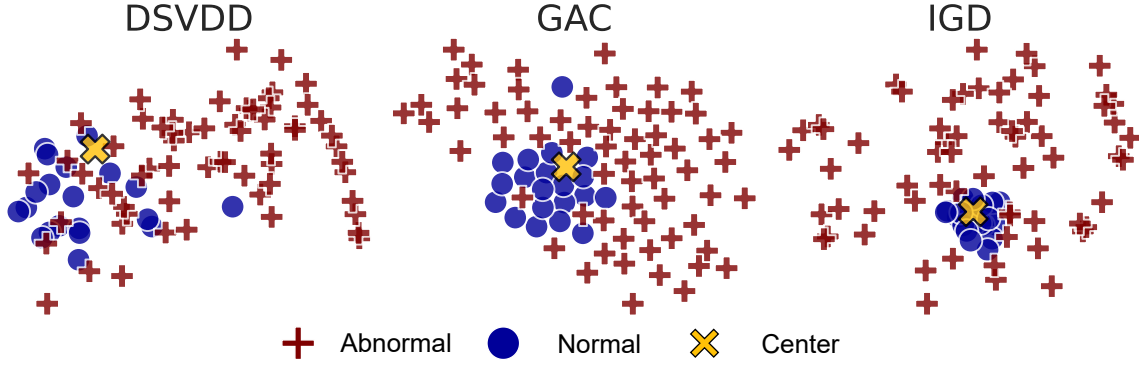


Figure A.2: t-sne visualisation from MVTeC (class bottle).

Metric	Method	Bottle	Hazelnut	Capsule	Metal Nut	Leather	Pill	Wood	Carpet	Tile	Grid	Cable	Transistor	Toothbrush	Screw	Zipper	Mean
Accuracy	AVID [195]	0.85	0.86	0.85	0.63	0.58	0.86	0.83	0.70	0.66	0.59	0.64	0.58	0.73	0.66	0.84	0.73
	AEssim [15]	0.88	0.54	0.61	0.54	0.46	0.60	0.83	0.67	0.52	0.69	0.61	0.52	0.74	0.51	0.80	0.63
	DAE [85]	0.80	0.88	0.62	0.73	0.44	0.62	0.74	0.50	0.77	0.78	0.56	0.71	0.98	0.69	0.80	0.71
	AnoGAN [200]	0.69	0.50	0.58	0.50	0.52	0.62	0.68	0.49	0.51	0.51	0.53	0.67	0.57	0.35	0.59	0.55
	λ -VAE _a [44]	0.86	0.74	0.86	0.78	0.71	0.80	0.89	0.67	0.81	0.83	0.56	0.70	0.89	0.71	0.67	0.77
	LSA [1]	0.86	0.80	0.71	0.67	0.70	0.85	0.75	0.74	0.70	0.54	0.61	0.50	0.89	0.75	0.88	0.73
	CAVGA-D _a [229]	0.89	0.84	0.83	0.67	0.71	0.88	0.85	0.73	0.70	0.75	0.63	0.73	0.91	0.77	0.87	0.78
	CAVGA-R _a [229]	0.91	0.87	0.87	0.71	0.75	0.91	0.88	0.78	0.72	0.78	0.67	0.75	0.97	0.78	0.94	0.82
	Ours - ImageNet	0.95	0.93	0.80	0.82	0.87	0.77	0.94	0.69	0.90	0.92	0.73	0.88	0.98	0.58	0.85	0.84
	Ours - SSL	0.95	0.93	0.81	0.82	0.90	0.74	0.89	0.71	0.94	0.90	0.79	0.85	0.98	0.67	0.88	0.85
AUC	AnoGAN [200]	0.800	0.259	0.442	0.284	0.451	0.711	0.567	0.337	0.401	0.871	0.477	0.692	0.439	0.100	0.715	0.503
	GANomaly [3]	0.794	0.874	0.721	0.694	0.808	0.671	0.920	0.821	0.720	0.743	0.711	0.808	0.700	1.000	0.744	0.782
	Skip-GANomaly [4]	0.937	0.906	0.718	0.790	0.908	0.758	0.919	0.795	0.850	0.657	0.674	0.814	0.689	1.000	0.663	0.805
	U-Net [189]	0.863	0.996	0.673	0.676	0.870	0.781	0.958	0.774	0.964	0.857	0.636	0.674	0.811	1.000	0.750	0.819
	DAGAN [215]	0.983	1.000	0.687	0.815	0.944	0.768	0.979	0.903	0.961	0.867	0.665	0.794	0.950	1.000	0.781	0.873
	SCADN [251]	0.957	0.856	0.765	0.504	0.983	0.833	0.659	0.624	0.814	0.831	0.792	0.981	0.863	0.968	0.846	0.818
	Ours - ImageNet	1.000	0.986	0.907	0.886	0.922	0.870	0.982	0.828	0.979	0.979	0.856	0.909	0.997	0.815	0.969	0.926
	Ours - SSL	1.000	0.997	0.915	0.913	0.958	0.873	0.946	0.828	0.991	0.978	0.906	0.906	0.997	0.825	0.970	0.934

Table A.1: **Anomaly detection**: mean testing accuracy and AUC on MVTeC AD produced by the SOTA and our method.

A.6 Correctness Proof

Lemma A.6.1. *Assuming that the maximisation of the constrained ℓ_{ELBO} produces θ that makes*

$$\mathbb{E}_{q(\omega)}[\log p_{\theta}(y=0, \omega | \mathbf{x}, \mathcal{P}_{\mathcal{X}})] \geq \mathbb{E}_{q(\omega)}[\log p_{\theta^{old}}(y=0, \omega | \mathbf{x}, \mathcal{P}_{\mathcal{X}})],$$

we have that $(\log p_{\theta}(y=0 | \mathbf{x}, \mathcal{P}_{\mathcal{X}}) - \log p_{\theta^{old}}(y=0 | \mathbf{x}, \mathcal{P}_{\mathcal{X}}))$ is lower bounded by

$$(\mathbb{E}_{q(\omega)}[\log p_{\theta}(y=0, \omega | \mathbf{x}, \mathcal{P}_{\mathcal{X}})] - \mathbb{E}_{q(\omega)}[\log p_{\theta^{old}}(y=0, \omega | \mathbf{x}, \mathcal{P}_{\mathcal{X}})]) \geq 0,$$

with $q(\omega) = p_{\theta^{old}}(\omega | \mathcal{P}_{\mathcal{X}})$.

Proof. We follow the proof for Theorem 1 in [46]. From the main paper, we have

$$\begin{aligned} \log p_\theta(y = 0|\mathbf{x}, \mathcal{P}_\mathcal{X}) = \\ \ell_{ELBO}(q, \theta) + KL[q(\omega)||p_\theta(\omega|\mathcal{P}_\mathcal{X})], \end{aligned} \quad (\text{A.5})$$

where $q(\omega) = p_{\theta^{old}}(\omega|\mathcal{P}_\mathcal{X})$. Subtracting $\log p_\theta(y = 0|\mathbf{x}, \mathcal{P}_\mathcal{X})$ and $\log p_{\theta^{old}}(y = 0|\mathbf{x}, \mathcal{P}_\mathcal{X})$, we have

$$\begin{aligned} \log p_\theta(y = 0|\mathbf{x}) - \log p_{\theta^{old}}(y = 0|\mathbf{x}) = \\ \ell_{ELBO}(q, \theta) - \ell_{ELBO}(q, \theta^{old}) + \\ KL[q(\omega)||p_\theta(\omega|\mathcal{P}_\mathcal{X})] - KL[q(\omega)||p_{\theta^{old}}(\omega|\mathcal{P}_\mathcal{X})]. \end{aligned} \quad (\text{A.6})$$

Since $KL[q(\omega)||p_\theta(\omega|\mathcal{P}_\mathcal{X})] \geq KL[q(\omega)||p_{\theta^{old}}(\omega|\mathcal{P}_\mathcal{X})]$ and that $\ell_{ELBO}(q, \theta) - \ell_{ELBO}(q, \theta^{old}) = \mathbb{E}_{q(\omega)}[\log p_\theta(y = 0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})] - \mathbb{E}_{q(\omega)}[\log p_{\theta^{old}}(y = 0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})]$, we conclude that

$$\begin{aligned} \log p_\theta(y = 0|\mathbf{x}, \mathcal{P}_\mathcal{X}) - \log p_{\theta^{old}}(y = 0|\mathbf{x}, \mathcal{P}_\mathcal{X}) \geq \\ \mathbb{E}_{q(\omega)}[\log p_\theta(y = 0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})] - \\ \mathbb{E}_{q(\omega)}[\log p_{\theta^{old}}(y = 0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})] \geq 0 \end{aligned} \quad (\text{A.7})$$

because of the assumption in this Lemma. \square

A.7 Convergence Conditions Proof

Lemma A.7.1. *Assume that $\{\theta^{(e)}\}_{e=1}^{+\infty}$ denotes the sequence of trained model parameters from the constrained optimisation of ℓ_{ELBO} such that: 1) the sequence $\{\log p_{\theta^{(e)}}(y = 0|\mathbf{x}, \mathcal{P}_\mathcal{X})\}_{e=1}^{+\infty}$ is bounded above, and 2) $(\mathbb{E}_{q(\omega)}[\log p_{\theta^{(e+1)}}(y = 0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})] - \mathbb{E}_{q(\omega)}[\log p_{\theta^{(e)}}(y = 0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})]) \geq \xi (\theta^{(e+1)} - \theta^{(e)})^\top (\theta^{(e+1)} - \theta^{(e)})$, for $\xi > 0$ and all $e \geq 1$, and $q(\omega) = p_{\theta^{(e)}}(\omega|\mathcal{P}_\mathcal{X})$. Then $\{\theta^{(e)}\}_{e=1}^{+\infty}$ converges to some $\theta^* \in \Theta$.*

Proof. We follow the proof for Theorem 2 in [46]. The sequence $\{\log p_{\theta^{(e)}}(y = 0|\mathbf{x}, \mathcal{P}_\mathcal{X})\}_{e=1}^{+\infty}$ is non-decreasing (from Lemma A.6.1) and bounded above (from assumption (1) in Lemma A.7.1), so it converges to $L^* < +\infty$. Hence, using Cauchy criterion [158], for any $\epsilon > 0$, we have $e^{(\epsilon)}$ such that, for $e \geq e^{(\epsilon)}$ and all $r \geq 1$,

$$\begin{aligned} \sum_{j=1}^r (\log p_{\theta^{(e+j)}}(y = 0|\mathbf{x}, \mathcal{P}_\mathcal{X}) - \log p_{\theta^{(e+j-1)}}(y = 0|\mathbf{x}, \mathcal{P}_\mathcal{X})) = \\ (\log p_{\theta^{(e+r)}}(y = 0|\mathbf{x}, \mathcal{P}_\mathcal{X}) - \log p_{\theta^{(e)}}(y = 0|\mathbf{x}, \mathcal{P}_\mathcal{X})) < \epsilon. \end{aligned} \quad (\text{A.8})$$

From (A.7),

$$\begin{aligned} 0 \leq \mathbb{E}_{q(\omega)}[\log p_{\theta^{(e+j)}}(y = 0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})] - \\ \mathbb{E}_{q(\omega)}[\log p_{\theta^{(e+j-1)}}(y = 0, \omega|\mathbf{x}, \mathcal{P}_\mathcal{X})] \\ \leq \log p_{\theta^{(e+j)}}(y = 0|\mathbf{x}, \mathcal{P}_\mathcal{X}) - \log p_{\theta^{(e+j-1)}}(y = 0|\mathbf{x}, \mathcal{P}_\mathcal{X}) \end{aligned} \quad (\text{A.9})$$

for $j \geq 1$ and $q(\omega) = p_{\theta^{(e+j-1)}}(\omega|\mathcal{P}_{\mathcal{X}})$. Hence, from (A.8),

$$\begin{aligned} & \sum_{j=1}^r (\mathbb{E}_{q(\omega)}[\log p_{\theta^{(e+j)}}(y=0, \omega|\mathbf{x}, \mathcal{P}_{\mathcal{X}})]) - \\ & \mathbb{E}_{q(\omega)}[\log p_{\theta^{(e+j-1)}}(y=0, z|\mathbf{x}, \mathcal{P}_{\mathcal{X}})]) < \epsilon, \end{aligned} \quad (\text{A.10})$$

for $e \geq e^{(\epsilon)}$ and all $r \geq 1$. Given assumption (2) in Lemma A.7.1 for $e, e+1, e+2, \dots, e+r-1$, we have from (A.10),

$$\epsilon > \xi \sum_{j=1}^r (\theta^{(e+j)} - \theta^{(e+j-1)})^\top (\theta^{(e+j)} - \theta^{(e+j-1)}), \quad (\text{A.11})$$

so

$$\epsilon > \xi (\theta^{(e+r)} - \theta^{(e)})^\top (\theta^{(e+r)} - \theta^{(e)}), \quad (\text{A.12})$$

which is a requirement to prove the convergence of $\theta^{(e)}$ to some $\theta^* \in \Theta$. \square

A.8 Class-level Results

The class-level results are shown in Tables A.1, A.2, A.3, A.4, and A.5. The mean accuracy and class-level anomaly detection accuracy on MVTEC dataset is displayed in Tab. A.1, where our ImageNet KD pre-trained model outperforms the previous SOTA methods CAVGA-D_u and CAVGA-R_u [228] by 6% and 2%, respectively, and our SSL pre-trained model outperforms their approach by 7% and 3%, respectively. With ImageNet KD pre-training, our model achieves the best accuracy results in **ten categories** of the MVTEC AD. The shallow generative baselines, such as DAE, AE-SSIM and AnoGAN yield sub-optimal results on MVTEC AD. When compared with methods recently considered to be the MVTEC AD SOTA, such as LSA [1] and λ -VAE_u [44], our approach shows more than 7% improvement. We also show the AUC anomaly detection results in Tab. A.1, where our method, with SSL and ImageNet KD pre-training, surpasses all previous methods by at least 5.3%, and produces the best results in eleven categories. The results of IGD for MNIST in Tab. A.2 show that our approach pre-trained with ImageNet KD is competitive with the Student-Teacher [11], and both are better than any of the previously proposed methods in the field. In Table A.3, we only show the results of our approach because we could not find the class-level results for other approaches. On the class-level results for CIFAR10, on Tab. A.4, we notice that our approach pre-trained with ImageNet and SSL shows the best AUC result in the field by a large margin (around 10%) compared with the Student-Teacher [11] approach. Finally, the class-level anomaly localisation AUC results for MVTEC on Tab. A.5 only shows the results of our approach because we could not find results from other approaches.

Method	0	1	2	3	4	5	6	7	8	9	Mean
DAE [85]	0.894	0.999	0.792	0.851	0.888	0.819	0.944	0.922	0.740	0.917	0.8766
VAE [109]	0.997	0.999	0.936	0.959	0.973	0.964	0.993	0.976	0.923	0.976	0.9696
KDE [19]	0.885	0.996	0.710	0.693	0.844	0.776	0.861	0.884	0.669	0.825	0.8140
OCSVM [201]	0.988	0.999	0.902	0.950	0.955	0.968	0.978	0.965	0.853	0.955	0.9510
AnoGAN [200]	0.966	0.992	0.850	0.887	0.894	0.883	0.947	0.935	0.849	0.924	0.9127
DSVDD [191]	0.980	0.997	0.917	0.919	0.949	0.885	0.983	0.946	0.939	0.965	0.9480
OCGAN [176]	0.998	0.999	0.942	0.963	0.975	0.980	0.991	0.981	0.939	0.981	0.9750
PixelCNN [224]	0.531	0.995	0.476	0.517	0.739	0.542	0.592	0.789	0.340	0.662	0.6180
CapsNet _{PP} [125]	0.998	0.990	0.984	0.976	0.935	0.970	0.942	0.987	0.993	0.990	0.9770
CapsNet _{RE} [125]	0.947	0.907	0.970	0.949	0.872	0.966	0.909	0.934	0.929	0.871	0.9250
ADGAN [41]	0.999	0.992	0.968	0.953	0.960	0.955	0.980	0.950	0.959	0.965	0.9680
LSA [1]	0.993	0.999	0.959	0.966	0.956	0.964	0.994	0.980	0.953	0.981	0.9750
GradCon [114]	0.995	0.999	0.952	0.973	0.969	0.977	0.994	0.979	0.919	0.973	0.9730
λ -VAE _u [44]	0.991	0.996	0.983	0.978	0.976	0.972	0.993	0.981	0.98	0.967	0.9820
ULSLM [243]	0.991	0.972	0.919	0.943	0.942	0.872	0.988	0.939	0.96	0.967	0.9490
CAVGA-D _u [228]	0.994	0.997	0.989	0.983	0.977	0.968	0.988	0.986	0.988	0.991	0.9860
Student-Teacher [11]	0.999	0.999	0.990	0.993	0.992	0.993	0.997	0.995	0.986	0.991	0.9935
Ours - ImageNet	0.998	0.999	0.992	0.991	0.993	0.991	0.997	0.990	0.984	0.991	0.9927

Table A.2: **Anomaly detection:** class-level testing AUC on MNIST produced by the SOTA and our methods.

A.9 Qualitative Localisation Results

Figure A.3 shows the polyp segmentation results on Hyper-Kvasir testing set images, and Figure A.4 displays the defect results on MVTEC AD testing set images.

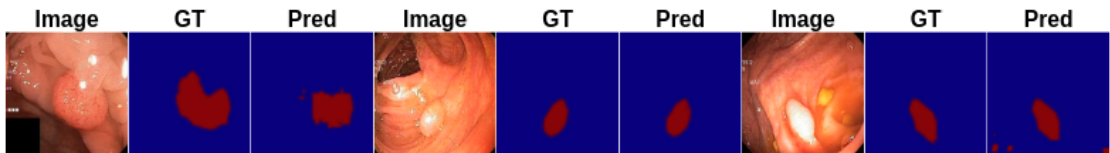


Figure A.3: Qualitative visual results from Hyper-Kvasir testing set (red = anomaly).

Method	0	1	2	3	4	5	6	7	8	9	Mean
Ours - ImageNet	0.908	0.992	0.902	0.946	0.93	0.95	0.818	0.993	0.938	0.981	0.935
Ours - SSL	0.926	0.992	0.922	0.946	0.931	0.971	0.832	0.992	0.946	0.982	0.944

Table A.3: **Anomaly detection:** class-level testing AUC on FMNIST produced by our methods.

Method	Plane	Car	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	Mean
DAE [85]	0.411	0.478	0.616	0.562	0.728	0.513	0.688	0.497	0.487	0.378	0.5358
VAE [109]	0.634	0.442	0.640	0.497	0.743	0.515	0.745	0.527	0.674	0.416	0.5833
KDE [19]	0.658	0.520	0.657	0.497	0.727	0.496	0.758	0.564	0.680	0.540	0.6100
OCSVM [201]	0.630	0.440	0.649	0.487	0.735	0.500	0.725	0.533	0.649	0.508	0.5860
AnoGAN [200]	0.671	0.547	0.529	0.545	0.651	0.603	0.585	0.625	0.758	0.665	0.6179
DSVDD [191]	0.617	0.659	0.508	0.591	0.609	0.657	0.677	0.673	0.759	0.731	0.6481
OCGAN [176]	0.757	0.531	0.640	0.62	0.723	0.620	0.723	0.575	0.820	0.554	0.6566
PixelCNN [224]	0.788	0.428	0.617	0.574	0.511	0.571	0.422	0.454	0.715	0.426	0.5510
CapsNet _{PP} [125]	0.622	0.455	0.671	0.675	0.683	0.350	0.727	0.673	0.710	0.466	0.6120
CapsNet _{RE} [125]	0.371	0.737	0.421	0.588	0.388	0.601	0.491	0.631	0.410	0.671	0.5310
ADGAN [41]	0.671	0.547	0.529	0.545	0.651	0.603	0.585	0.625	0.758	0.665	0.6180
LSA [1]	0.735	0.580	0.690	0.542	0.761	0.546	0.751	0.535	0.717	0.548	0.6410
GradCon [114]	0.760	0.598	0.648	0.586	0.733	0.603	0.684	0.567	0.784	0.678	0.6640
λ -VAE _u [44]	0.702	0.663	0.68	0.713	0.77	0.689	0.805	0.588	0.813	0.744	0.7170
ULSLM [243]	0.740	0.747	0.628	0.572	0.678	0.602	0.753	0.685	0.781	0.795	0.7360
CAVGA-D _u [228]	0.653	0.784	0.761	0.747	0.775	0.552	0.813	0.745	0.701	0.741	0.7370
Student-Teacher [11]	0.789	0.849	0.734	0.748	0.851	0.793	0.892	0.830	0.862	0.848	0.8196
Ours - ImageNet	0.868	0.870	0.738	0.716	0.850	0.766	0.890	0.871	0.898	0.899	0.8368
Ours - SSL	0.906	0.979	0.839	0.823	0.886	0.899	0.909	0.964	0.969	0.948	0.9125

Table A.4: **Anomaly detection:** class-level testing AUC on CIFAR10 produced by the SOTA and our methods.

Method	Bottle	Hazelnut	Capsule	Metal Nut	Leather	Pill	Wood	Carpet	Tile	Grid	Cable	Transistor	Toothbrush	Screw	Zipper	Mean
Ours - ImageNet	0.928	0.981	0.967	0.902	0.983	0.962	0.827	0.901	0.727	0.916	0.835	0.843	0.974	0.960	0.932	0.909
Ours - SSL	0.922	0.980	0.977	0.926	0.995	0.973	0.891	0.947	0.780	0.977	0.847	0.844	0.977	0.970	0.967	0.931

Table A.5: **Anomaly localisation:** class-level testing pixel-wise localisation AUC results on the anomalous images of MVTec AD produced by our methods.

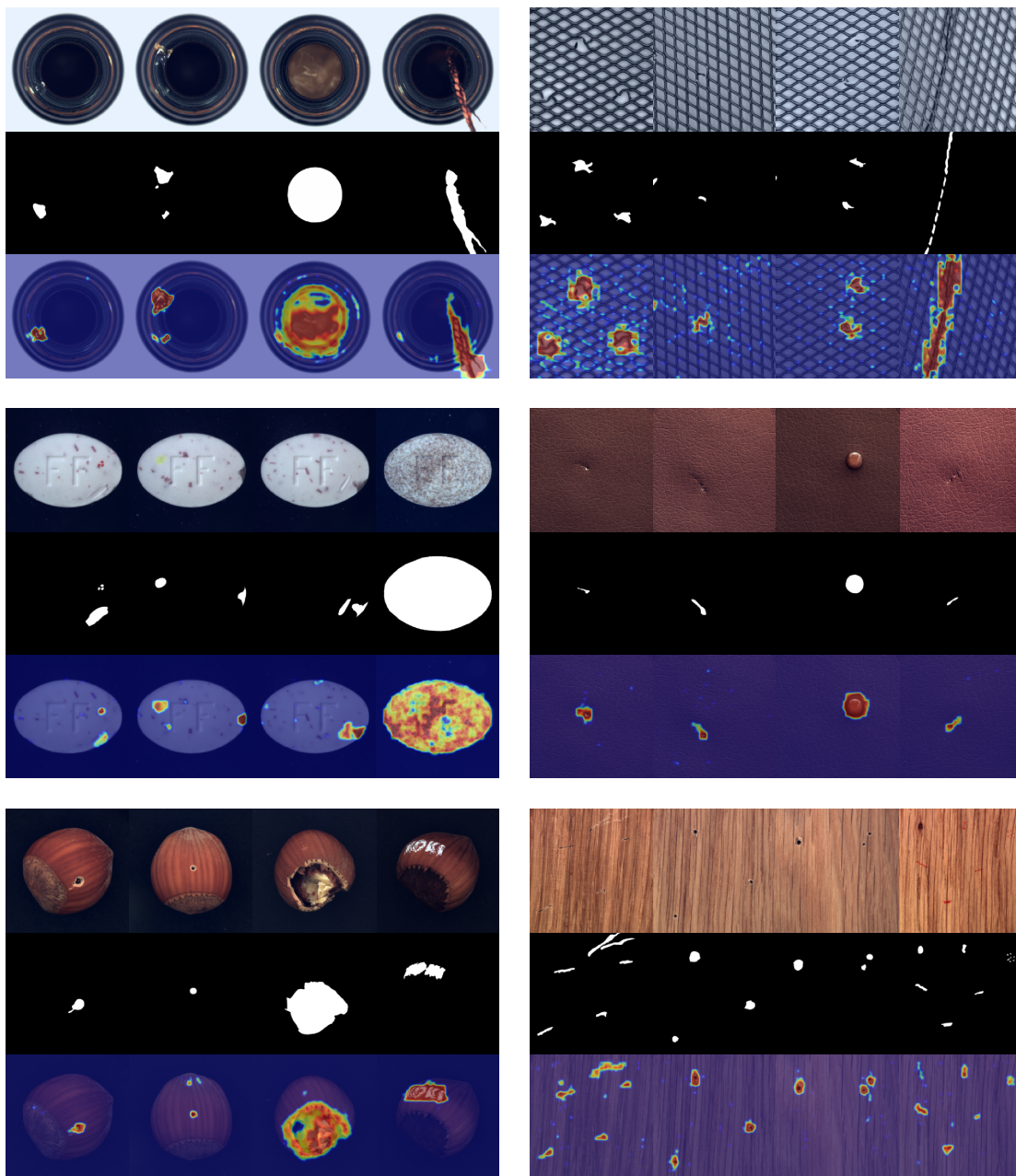


Figure A.4: Qualitative results of our anomaly localisation results on the MVtec AD testing set (red = high probability of anomaly).

Appendix B

RTFM (Chapter 8) Appendix

B.1 Theoretical Motivation of RTFM

Theorem B.1.1 (Expected Separability Between Abnormal and Normal Videos). *Assuming that $\mathbb{E}[\|\mathbf{x}^+\|_2] \geq \mathbb{E}[\|\mathbf{x}^-\|_2]$, where \mathbf{X}^+ has μ abnormal samples and $(T - \mu)$ normal samples, where $\mu \in [1, T]$, and \mathbf{X}^- has T normal samples. Let $D_{\theta,k}(\cdot)$ be the random variable from which the separability scores $d_{\theta,k}(\cdot)$ of Eq.3 in the main paper are drawn [124].*

1. If $0 < k < \mu$, then

$$0 \leq \mathbb{E}[D_{\theta,k}(\mathbf{X}^+, \mathbf{X}^-)] \leq \mathbb{E}[D_{\theta,k+1}(\mathbf{X}^+, \mathbf{X}^-)].$$

2. For a finite μ , then

$$\lim_{k \rightarrow \infty} \mathbb{E}[D_{\theta,k}(\mathbf{X}^+, \mathbf{X}^-)] = 0.$$

Proof.

$$\begin{aligned} \mathbb{E}[D_{\theta,k}(\mathbf{X}^+, \mathbf{X}^-)] &= \mathbb{E}[g_{\theta,k}(\mathbf{X}^+)] - \mathbb{E}[g_{\theta,k}(\mathbf{X}^-)] \\ &= p_k^+(\mathbf{X}^+) \mathbb{E}[\|\mathbf{x}^+\|_2] + p_k^-(\mathbf{X}^+) \mathbb{E}[\|\mathbf{x}^-\|_2] - \mathbb{E}[\|\mathbf{x}^-\|_2] \end{aligned} \tag{B.1}$$

1. Trivial given that $\mathbb{E}[\|\mathbf{x}^+\|_2] \geq \mathbb{E}[\|\mathbf{x}^-\|_2]$ and that $p_{k+1}^+(\mathbf{X}^+) > p_k^+(\mathbf{X}^+)$ for $0 < k < \mu$
2. Trivial given that as μ is finite, $\lim_{k \rightarrow \infty} p_k^+(\mathbf{X}^+) = 0$.

□

Intuition of feature magnitude: Assuming the expected magnitude of abnormal samples is larger than of normal samples, we can derive Thm. 3.1 that proves that the expected feature magnitude-based separability score between normal and abnormal videos grows for $0 < k < \mu$ and reduces to zero for $k \rightarrow \infty$. Hence, to use Thm. 3.1, we need to enforce larger magnitude for abnormal features using our proposed RTFM. The similarity between the theoretical and empirical curves in Fig.B.1(left) is evidence of the soundness of Thm. 3.1.

B.2 Computational Efficiency

We investigate if our system can run in real time. During inference, our method processes a 16-frame clip in 0.76 seconds on a Nvidia 2080Ti—this time includes the I3D extraction time. This indicates that our system can achieve good real-time detection in real-world applications.

B.3 Temporal Dependency

Temporal Dependency has been explored in [65, 112, 137, 145, 245, 250, 266]. In anomaly detection, traditional methods [112, 250] convert consecutive frames into handcrafted motion trajectories to capture the local consistency between neighbouring frames. Diverse temporal dependency modelling methods have been used in deep anomaly detection approaches, such as stacked RNN [145], temporal consistency in future frame prediction [65], and convolution LSTM [137]. However, these methods capture short-range fixed-order temporal correlations only with single temporal scale, ignoring the long-range dependency from all possible temporal locations and the events with varying temporal length. GCN-based methods are explored in [245, 266] to capture the long-range dependency from snippets features, but they are inefficient and hard to train. By contrast, our proposed module combines PDC [254] and TSA [239] on the temporal dimension to seamlessly and efficiently incorporate both the long and short-range temporal dependencies into our temporal feature ranking loss.

B.4 Ablations for k and m

We show the AUC results as a function of top- k and margin m values on ShanghaiTech in Fig.B.1. Consistent to our theoretical analysis, the performance of our model peaks at a sufficiently large k , flattens at around $k \approx \mu$ and then drops with increasing k (Fig.B.1(left)). It is also robust to a large range of $m \in [50, 1200]$ with a stable AUC in [93%, 96%] (Fig.B.1(right)).

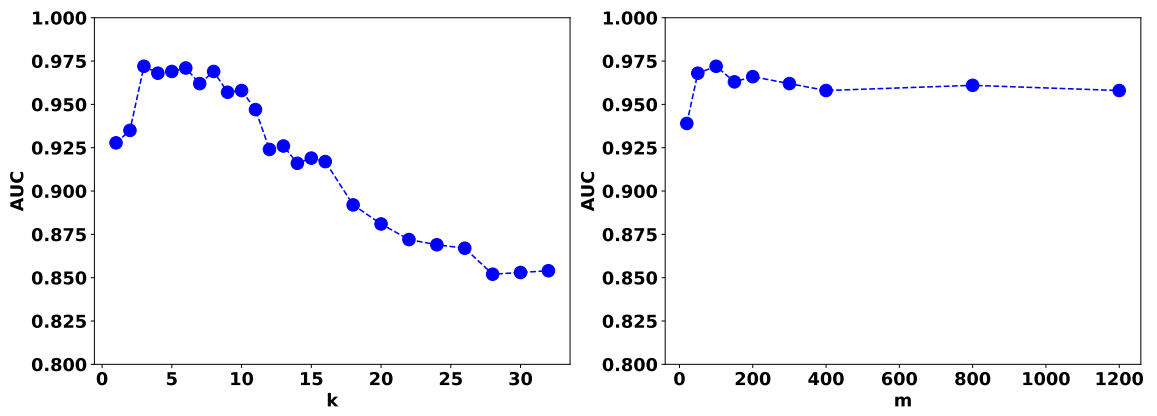


Figure B.1: AUC w.r.t. top- k (Left) and the margin m (Right).

Appendix C

PEBAL (Chapter 11) Appendix

C.1 Qualitative results

In Figure C.1, we show some additional qualitative results. Our approach can effectively detect small and distant objects (rows 6 and 7) and objects with different scales (rows 1 to 5).

C.2 More AUC results

In Tables C.1 and C.2, we show the AUC results in addition to the AP and FPR results in Tables 6 and 7 of the main paper. We achieve consistently SOTA AUC performance regardless of the selection of outlier classes or the number of outlier training samples.

Class Per.	FS LF - AUC	FS Static - AUC
1%	97.59 \pm 0.39	98.37 \pm 0.56
5%	98.17 \pm 0.45	98.25 \pm 0.71
10%	98.47 \pm 0.39	99.59 \pm 0.25
25%	98.39 \pm 0.28	99.52 \pm 0.17
50%	98.63 \pm 0.07	99.54 \pm 0.08
75%	98.71 \pm 0.05	99.59 \pm 0.03

Table C.1: AUC testing results (mean results over six random seeds) of our approach on Fishyscapes benchmark w.r.t. different **diversity of OE classes**.

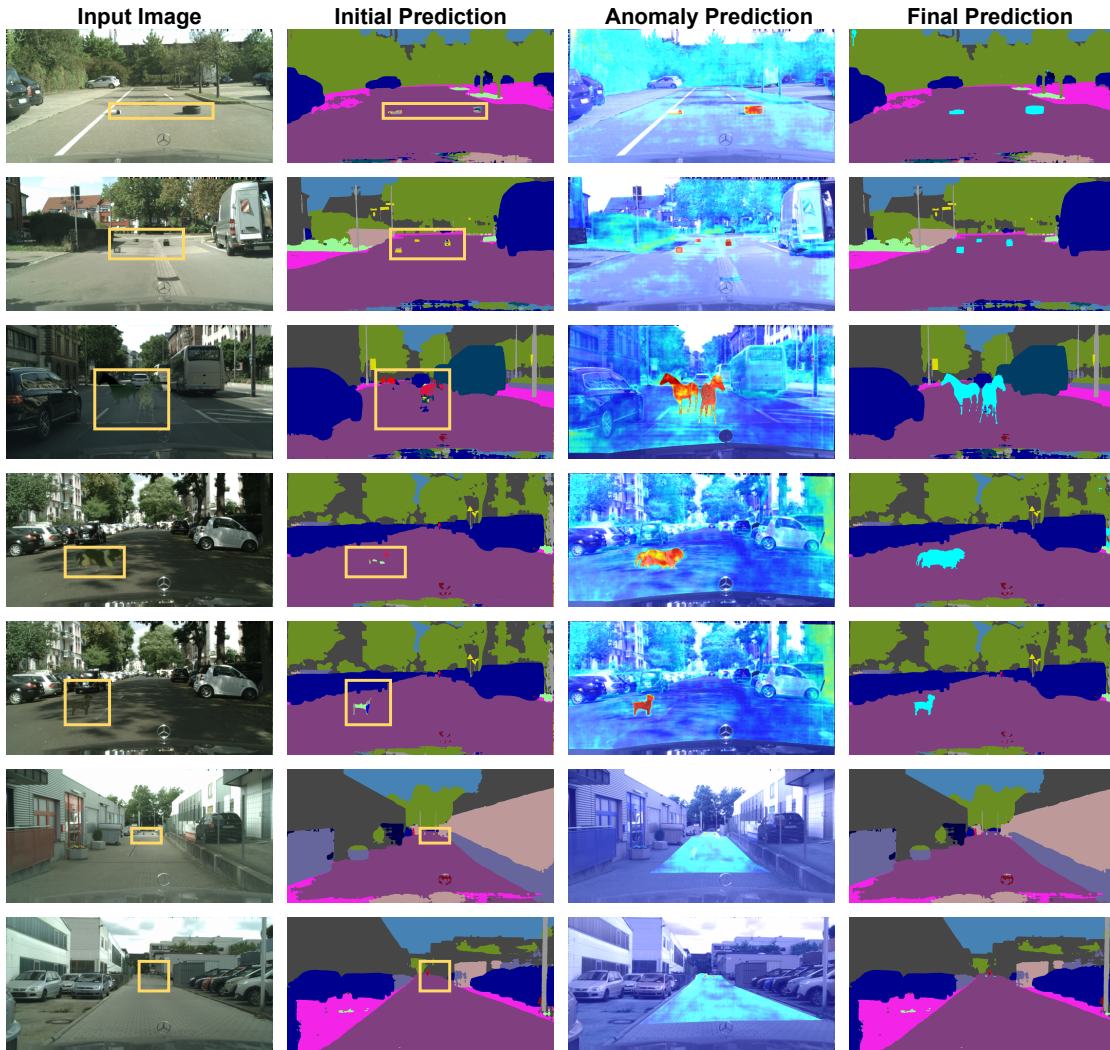


Figure C.1: From the **input image** (anomaly highlighted with a yellow box), the **initial prediction** shows the original segmentation results with anomalies classified as a one of the pre-defined inlier classes. **Anomaly predictions** from **our** method show an anomaly map with high scores (in yellow and red) for anomalous pixels. In our **final prediction**, anomalous pixels are coloured in cyan.

C.3 Hyper-parameters Selection

For testing, we note a small performance gap with $\lambda \in \{0.1, 0.01\}$ on LF test set, with AP=78.29 for $\lambda = 0.01$ and AP=77.15 for $\lambda = 0.1$. For the EBM margin, PEBAL reaches AP $\in [76.9, 78.3]$ and FPR $\in [0.8, 1.3]$ for $m_{in} \in [-12, -22]$ and $m_{out} \in [-2, -8]$

Train Size	FS LF - AUC	FS Static - AUC
5%	98.13 \pm 0.12	99.16 \pm 0.09
10%	98.35 \pm 0.15	99.57 \pm 0.07
25%	98.36 \pm 0.06	99.51 \pm 0.06
50%	98.69 \pm 0.05	99.37 \pm 0.07

Table C.2: AUC testing results (mean results over six random seeds) of our approach on Fishyscapes benchmark w.r.t. different **amount of OE training samples**.

for different values of m_{in} and m_{out} on LF test set.

C.4 Training Details on Cityscapes

Following [26, 27], we use the same DeepLabv3+ [28] with WideResNet34 (90.3 mIoU on Cityscapes Val) trained by Nvidia [271] as one of the backbones of our segmentation model. As mentioned in [271], the model is firstly pre-trained on Mapillary Vista dataset [156], and then fine-tuned on Cityscapes train set with their proposed label relaxation loss and sdc-aug label propagation. Their model uses a different {cv2: monchengladbach, strasbourg, stuttgart} validation split than the standard split {cv0: munster, lindau, frankfurt}. Please refer to their paper for more details. For DeepLabv3+ [28] with Resnet101 backbone (80.3 mIoU on Cityscapes Val) from [103], the authors trained their model with the standard cv0 train/validation split using default formulations in [28]. All those checkpoints are downloaded from their official Github pages.

C.5 Results Based on Different DeepLabv3+ Checkpoint

In this section, we show the results of another DeepLabv3+ [28] with WideResNet34 trained by Nvidia [271] using the Cityscapes **{cv0: munster, lindau, frankfurt}** standard train/val split. The checkpoint is downloaded from the their official Github page [271], with a 81.8% mIoU on Cityscapes validation set. This model was firstly pre-trained on Mapillary Vista dataset [156] and then fine-tuned on Cityscapes but without their label relaxation loss and sdc-aug label propagation. As shown in Tab. C.3, our model outperforms the previous methods by a large margin on all three benchmarks, regardless of the backbones, the segmentation accuracy and the Cityscapes train/val

Table C.3: Anomaly segmentation results on **Fishyscapes validation sets** (LostAnd-Found and Static), and the **Road Anomaly testing set**, with **WideResnet34** backbone under **cv0** standard train/val split.

Methods	FS LostAndFound			FS Static			Road Anomaly		
	AUC \uparrow	AP \uparrow	FPR ₉₅ \downarrow	AUC \uparrow	AP \uparrow	FPR ₉₅ \downarrow	AUC \uparrow	AP \uparrow	FPR ₉₅ \downarrow
MSP [93]	89.26	11.84	32.55	89.26	11.84	32.55	72.37	20.23	67.98
Max Logit [93]	93.14	12.78	38.15	93.27	18.89	25.49	76.39	23.46	64.55
Entropy [95]	89.01	8.79	47.81	90.28	15.19	31.71	73.70	22.13	67.42
Energy [138]	93.45	14.29	37.71	93.52	19.22	25.02	76.76	23.48	64.04
SML [103]	96.03	21.71	20.09	95.79	32.04	15.81	74.45	22.16	68.59
Ours	98.52	64.43	6.56	99.33	86.01	2.63	88.85	44.41	37.98

splits. Notably, our method surpasses the previous SOTA SML by 40%, 50% and 20% of AP on three datasets, respectively. We also achieve best AUC and FPR results on all datasets.

Bibliography

- [1] Davide Abati et al. Latent space autoregression for novelty detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 481–490, 2019. [15](#), [85](#), [87](#), [88](#), [176](#), [178](#), [179](#), [180](#)
- [2] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. [12](#), [110](#)
- [3] Samet Akçay et al. Ganomaly: Semi-supervised anomaly detection via adversarial training. In Asian conference on computer vision, pages 622–637. Springer, 2018. [50](#), [87](#), [176](#)
- [4] Samet Akçay et al. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In 2019 International Joint Conference on Neural Networks (IJCNN), pages 1–8, 2019. [87](#), [176](#)
- [5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. arXiv preprint arXiv:1701.07875, 2017. [12](#), [23](#), [24](#)
- [6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6836–6846, 2021. [132](#)
- [7] Arslan Basharat, Alexei Gritai, and Mubarak Shah. Learning object motion patterns for anomaly detection and improved object detection. In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2008. [110](#)
- [8] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In International MICCAI Brainlesion Workshop, pages 161–169. Springer, 2018. [12](#), [14](#), [76](#), [156](#)
- [9] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Scale-space autoencoders for unsupervised anomaly segmentation in brain mri. In MICCAI, pages 552–561. Springer, 2020. [34](#), [48](#), [96](#)
- [10] Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. arXiv preprint arXiv:2005.02359, 2020. [4](#), [11](#), [34](#), [35](#), [36](#), [48](#), [50](#)

- 75, 76, 85, 86, 110, 170
- [11] Paul Bergmann et al. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4183–4192, 2020. 12, 15, 58, 75, 76, 84, 85, 86, 178, 179, 180
 - [12] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 110
 - [13] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9592–9600, 2019. xvii, xx, 12, 15, 73, 76, 82, 84, 173, 174
 - [14] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020. 110
 - [15] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. arXiv preprint arXiv:1807.02011, 2018. 12, 75, 76, 87, 88, 176
 - [16] David Berthelot, Colin Raffel, Aurko Roy, and Ian Goodfellow. Understanding and improving interpolation in autoencoders via an adversarial regularizer. arXiv preprint arXiv:1807.07543, 2018. 38, 56, 79
 - [17] Petra Bevandi et al. Simultaneous semantic segmentation and outlier detection in presence of domain shift. In German Conference on Pattern Recognition, pages 33–47. Springer, 2019. 5, 14, 154, 157, 163, 164
 - [18] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Discriminative out-of-distribution detection for semantic segmentation. arXiv preprint arXiv:1808.07703, 2018. 14, 15, 157
 - [19] Christopher M Bishop. Pattern recognition and machine learning. springer, 2006. 85, 179, 180
 - [20] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. arXiv preprint arXiv:1904.03215, 2019. 5, 154, 156, 157, 161, 162, 164
 - [21] Hanna Borgli and et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific Data, 7(1):1–14, 2020. xv, xvi, xviii, 15, 17, 35, 38, 39, 42, 49, 56, 57, 62, 65, 66, 84, 97, 101, 102, 103, 135
 - [22] Philippe Burlina, Neil Joshi, and I-Jeng Wang. Where’s wally now? deep gener-

- ative and discriminative embeddings for novelty detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. [110](#)
- [23] Saskia Camps, Tim Houben, Davide Fontanarosa, Christopher Edwards, Maria Antico, Matteo Dunnhofer, Esther Martens, Jose Baeza, Ben Vanneste, Evert van Limbergen, et al. One-class gaussian process regressor for quality assessment of transperineal ultrasound images. In Proceedings of the 1st International Conference on Medical Imaging with Deep Learning 2018, pages 1–10. Medical Imaging with Deep Learning Conference Committee, 2018. [11](#), [22](#), [24](#)
- [24] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308, 2017. [xix](#), [111](#), [131](#), [132](#)
- [25] Bing-Bing Chai, Jozsef Vass, and Xinhua Zhuang. Significance-linked connected component analysis for wavelet image coding. IEEE Transactions on Image processing, 8(6):774–784, 1999. [42](#)
- [26] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Mathieu Salzmann, Pascal Fua, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation. NeurIPS, 2021. [162](#), [189](#)
- [27] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5128–5137, 2021. [xx](#), [5](#), [14](#), [15](#), [154](#), [155](#), [157](#), [158](#), [160](#), [162](#), [163](#), [167](#), [189](#)
- [28] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV), pages 801–818, 2018. [162](#), [189](#)
- [29] Ting Chen et al. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020. [85](#), [175](#)
- [30] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In ICML, pages 1597–1607. PMLR, 2020. [4](#), [34](#), [35](#), [36](#), [39](#), [40](#), [41](#), [170](#)
- [31] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In ICML, pages 1597–1607. PMLR, 2020. [11](#), [12](#), [48](#), [50](#), [57](#), [68](#), [69](#)
- [32] Xiaoran Chen, Suhang You, Kerem Can Tezcan, and Ender Konukoglu. Unsupervised lesion detection via image restoration with a normative prior. Medical image analysis, 64:101713, 2020. [4](#), [34](#), [48](#), [50](#)
- [33] Yuanhong Chen, Yu Tian, Guansong Pang, and Gustavo Carneiro. Deep

- one-class classification via interpolated gaussian descriptor. arXiv preprint arXiv:2101.10043, 2021. [2](#), [14](#), [96](#), [100](#), [101](#), [102](#), [103](#), [104](#), [156](#)
- [34] Yuanhong Chen, Yu Tian, Guansong Pang, and Gustavo Carneiro. Unsupervised anomaly detection and localisation with multi-scale interpolated gaussian descriptors. arXiv preprint arXiv:2101.10043, 2021. [ix](#), [xvi](#), [xvii](#), [4](#), [34](#), [35](#), [37](#), [38](#), [39](#), [41](#), [42](#), [48](#), [49](#), [50](#), [55](#), [56](#), [58](#), [59](#), [60](#), [61](#), [62](#), [64](#), [66](#), [67](#), [110](#)
- [35] Yunqiang Chen, Xiang Sean Zhou, and Thomas S Huang. One-class svm for learning in image retrieval. In Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205), volume 1, pages 34–37. IEEE, 2001. [1](#), [50](#)
- [36] Kai-Wen Cheng, Yie-Tarng Chen, and Wen-Hsien Fang. Video anomaly detection and localization using hierarchical feature representation and gaussian process regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015. [110](#)
- [37] Hyunsoo Cho, Jinseok Seol, and Sang-goo Lee. Masked contrastive learning for anomaly detection. arXiv preprint arXiv:2105.08793, 2021. [53](#)
- [38] François Chollet. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1251–1258, 2017. [xix](#), [132](#)
- [39] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3213–3223, 2016. [5](#), [156](#), [162](#)
- [40] Clement Creusot and Asim Munawar. Real-time small obstacle detection on highways using compressive rbm road reconstruction. In 2015 IEEE Intelligent Vehicles Symposium (IV), pages 162–167. IEEE, 2015. [14](#), [156](#)
- [41] Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft. Image anomaly detection with generative adversarial networks. In Joint european conference on machine learning and knowledge discovery in databases, pages 3–17. Springer, 2018. [15](#), [85](#), [179](#), [180](#)
- [42] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. arXiv preprint arXiv:2011.08785, 2020. [11](#)
- [43] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In International Conference on Pattern Recognition, pages 475–489. Springer, 2021. [x](#), [xvii](#), [49](#), [55](#), [56](#), [58](#), [59](#), [60](#), [61](#), [62](#), [64](#), [67](#), [68](#), [75](#)
- [44] David Dehaene, Oriel Frigo, Sébastien Combexelle, and Pierre Eline. Iterative energy-based projection on a normal data manifold for anomaly localization. arXiv preprint arXiv:2002.03734, 2020. [85](#), [87](#), [88](#), [176](#), [178](#), [179](#), [180](#)

- [45] Allison Del Giorno, J Andrew Bagnell, and Martial Hebert. A discriminative framework for anomaly detection in large videos. In European Conference on Computer Vision, pages 334–349. Springer, 2016. [110](#)
- [46] Arthur P Dempster and Others. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22, 1977. [77](#), [80](#), [81](#), [177](#)
- [47] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 248–255. IEEE, 2009. [xvi](#), [50](#), [66](#)
- [48] Li Deng. The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 29(6):141–142, 2012. [15](#), [84](#)
- [49] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16918–16927, 2021. [2](#), [5](#), [14](#), [154](#), [156](#), [157](#), [161](#), [162](#), [163](#), [164](#), [165](#), [167](#)
- [50] Foivos I Diakogiannis et al. Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing, 162:94–114, 2020. [41](#), [42](#)
- [51] Carl Doersch. Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908, 2016. [28](#), [148](#)
- [52] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In ICCV, pages 1422–1430, 2015. [37](#), [54](#)
- [53] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. [99](#), [101](#), [132](#)
- [54] Ran El-Yaniv et al. On the foundations of noise-free selective classification. Journal of Machine Learning Research, 11(5), 2010. [157](#)
- [55] Alec Radford et al. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015. [12](#), [24](#), [27](#)
- [56] Dong Gong et al. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. arXiv preprint arXiv:1904.02639, 2019. [4](#), [12](#), [15](#), [23](#), [34](#), [48](#), [50](#), [75](#), [85](#), [96](#), [101](#), [110](#), [117](#), [119](#), [136](#), [142](#), [143](#)
- [57] Diederik P Kingma et al. Adam: A method for stochastic optimization, iclr, 2015. [27](#), [147](#)
- [58] Ian Goodfellow et al. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014. [4](#), [23](#), [34](#), [38](#)

- [59] Jorge Bernal et al. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. IEEE Transactions on Medical Imaging, 36(6):1231–1249, 2017. [22](#)
- [60] Jonathan Masci et al. Stacked convolutional auto-encoders for hierarchical feature extraction. In International Conference on Artificial Neural Networks, pages 52–59. Springer, 2011. [24](#), [28](#), [42](#), [59](#), [61](#), [89](#), [100](#), [101](#), [143](#), [148](#), [149](#)
- [61] Leonardo Zorron Cheng Tao Pu et al. Computer-aided diagnosis for characterising colorectal lesions: Interim results of a newly developed software. Gastrointestinal Endoscopy, 87(6):AB245, 2018. [22](#)
- [62] Rebecca Siegel et al. Colorectal cancer statistics, 2014. CA: a cancer journal for clinicians, 64(2):104–117, 2014. [21](#), [22](#), [143](#)
- [63] Tim Salimans et al. Improved techniques for training gans. In Advances in neural information processing systems, pages 2234–2242, 2016. [27](#)
- [64] Thomas Schlegl et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In International Conference on Information Processing in Medical Imaging, pages 146–157. Springer, 2017. [12](#), [24](#), [143](#)
- [65] Wen Liu et al. Future frame prediction for anomaly detection—a new baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6536–6545, 2018. [3](#), [12](#), [16](#), [24](#), [108](#), [109](#), [110](#), [111](#), [117](#), [118](#), [119](#), [143](#), [184](#)
- [66] Xiaofei He et al. Laplacian score for feature selection. In Advances in neural information processing systems, pages 507–514, 2006. [25](#), [144](#)
- [67] Yu Tian et al. One-stage five-class polyp detection and classification. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pages 70–73. IEEE, 2019. [21](#), [22](#), [34](#), [48](#), [96](#), [130](#), [143](#)
- [68] Yiru Zhao et al. Spatio-temporal autoencoder for video anomaly detection. In Proceedings of the 25th ACM international conference on Multimedia, pages 1933–1941. ACM, 2017. [24](#)
- [69] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2):303–338, 2010. [161](#)
- [70] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In MICCAI, pages 263–273. Springer, 2020. [34](#), [48](#), [96](#)
- [71] Yuqi Fang, Cheng Chen, Yixuan Yuan, and Kai-yu Tong. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In MICCAI, pages 302–310. Springer, 2019. [41](#), [42](#)
- [72] Zhiwen Fang, Jiafei Liang, Joey Tianyi Zhou, Yang Xiao, and Feng Yang. Anomaly detection with bidirectional consistency in videos. IEEE Transactions on Neural Networks and Learning Systems, 2020. [11](#), [110](#)

- [73] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14009–14018, 2021. [13](#), [130](#), [136](#)
- [74] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. arXiv preprint arXiv:1703.03400, 2017. [143](#)
- [75] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. arXiv preprint arXiv:2011.07491, 2020. [118](#)
- [76] Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 270–279, 2017. [84](#)
- [77] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. arXiv preprint arXiv:1805.10917, 2018. [ix](#), [4](#), [11](#), [34](#), [35](#), [36](#), [41](#), [42](#), [48](#), [50](#), [68](#), [69](#), [75](#), [76](#), [85](#), [86](#), [110](#), [170](#)
- [78] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016. [148](#)
- [79] Ian Goodfellow et al. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 27, pages 2672–2680. Curran Associates, Inc., 2014. [79](#), [80](#)
- [80] Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge distillation: A survey. arXiv preprint arXiv:2006.05525, 2020. [85](#)
- [81] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. arXiv preprint arXiv:1912.03263, 2019. [154](#), [157](#), [159](#)
- [82] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Dense anomaly detection by robust learning on synthetic negative data. arXiv preprint arXiv:2112.12833, 2021. [154](#)
- [83] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In Advances in neural information processing systems, pages 5767–5777, 2017. [26](#)
- [84] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. arXiv preprint arXiv:1706.04599, 2017. [167](#)
- [85] Raia Hadsell et al. Dimensionality reduction by learning an invariant mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, pages 1735–1742. IEEE, 2006. [85](#), [87](#), [88](#), [176](#), [179](#), [180](#)

- [86] David Haldimann, Hermann Blum, Roland Siegwart, and Cesar Cadena. This is not what i imagined: Error detection for semantic segmentation through visual dissimilarity. arXiv preprint arXiv:1909.00676, 2019. [14](#), [156](#)
- [87] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 733–742, 2016. [3](#), [108](#), [118](#), [119](#), [120](#)
- [88] Chengkun He, Jie Shao, and Jiayu Sun. An anomaly-introduced learning method for abnormal event detection. Multimedia Tools and Applications, 77(22):29573–29588, 2018. [17](#), [117](#)
- [89] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377, 2021. [96](#), [97](#), [101](#), [103](#), [169](#), [172](#)
- [90] Kaiming He et al. Momentum contrast for unsupervised visual representation learning. In CVPR, pages 9729–9738, 2020. [4](#), [11](#), [12](#), [34](#), [35](#), [36](#), [40](#), [41](#), [48](#), [50](#), [170](#)
- [91] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. [39](#), [57](#)
- [92] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. [175](#)
- [93] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. arXiv preprint arXiv:1911.11132, 2019. [157](#), [162](#), [163](#), [164](#), [165](#), [190](#)
- [94] Dan Hendrycks et al. Using self-supervised learning can improve model robustness and uncertainty. arXiv preprint arXiv:1906.12340, 2019. [4](#), [11](#), [34](#), [35](#), [48](#), [50](#), [170](#)
- [95] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136, 2016. [14](#), [156](#), [162](#), [163](#), [164](#), [165](#), [190](#)
- [96] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. arXiv preprint arXiv:1812.04606, 2018. [14](#), [154](#), [157](#)
- [97] Ryota Hinami, Tao Mei, and Shin’ichi Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In Proceedings of the IEEE International Conference on Computer Vision, pages 3619–3627, 2017. [3](#), [108](#)
- [98] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. arXiv preprint

- arXiv:1808.06670, 2018. [xix](#), [142](#), [144](#), [145](#), [146](#), [147](#)
- [99] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. [148](#)
- [100] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7842–7851, 2019. [75](#), [110](#)
- [101] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu. Unmasking the abnormal events in video. In Proceedings of the IEEE International Conference on Computer Vision, pages 2895–2903, 2017. [11](#), [110](#)
- [102] Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Huazhu Fu, Debesh Jha, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 142–152. Springer, 2021. [130](#)
- [103] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15425–15434, 2021. [5](#), [14](#), [154](#), [156](#), [157](#), [161](#), [162](#), [163](#), [164](#), [165](#), [189](#), [190](#)
- [104] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In CVPR, 2014. [118](#)
- [105] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017. [118](#), [135](#)
- [106] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In Advances in neural information processing systems, pages 5574–5584, 2017. [14](#), [156](#)
- [107] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. [101](#), [118](#), [135](#)
- [108] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013. [4](#), [34](#)
- [109] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013. [85](#), [179](#), [180](#)
- [110] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. CoRR, abs/1901.09005, 2019. [37](#)
- [111] Bruno Korbar, Andrea M Olofson, Allen P Mirafior, Catherine M Nicka, Matthew A Suriawinata, Lorenzo Torresani, Arief A Suriawinata, and Saeed Hassanpour. Deep learning for classification of colorectal polyps on whole-slide

- images. Journal of pathology informatics, 8, 2017. 143
- [112] Louis Kratz and Ko Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 1446–1453. IEEE, 2009. 111, 184
- [113] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The cifar-10 dataset. online: <http://www.cs.toronto.edu/kriz/cifar.html>, 55, 2014. 15, 84
- [114] Gukyeong Kwon, Mohit Prabhushankar, Dogancan Temel, and Ghassan AlRegib. Backpropagated gradient representations for anomaly detection, 2020. 85, 179, 180
- [115] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. arXiv preprint arXiv:1612.01474, 2016. 14, 156
- [116] Fahad Lateef and Yassine Ruichek. Survey on semantic segmentation using deep learning techniques. Neurocomputing, 338:321–348, 2019. 14, 153
- [117] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. Predicting structured data, 1(0), 2006. 154, 157, 159
- [118] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. arXiv preprint arXiv:1711.09325, 2017. 14, 156
- [119] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems, 31, 2018. 162, 163, 164
- [120] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9664–9674, 2021. 55, 59, 60, 61, 62, 75, 96, 160
- [121] Liu Li et al. Attention based glaucoma detection: A large-scale database and cnn model. In CVPR, pages 10571–10580, 2019. xvi, 15, 16, 35, 38, 39, 49, 56, 57, 62, 63, 84
- [122] Liu Li, Mai Xu, Xiaofei Wang, Lai Jiang, and Hanruo Liu. Attention based glaucoma detection: A large-scale database and cnn model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 12, 76
- [123] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. IEEE transactions on pattern analysis and machine intelligence, 36(1):18–32, 2013. 110
- [124] Weixin Li and Nuno Vasconcelos. Multiple instance learning for soft bags via top instances. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4277–4285, 2015. 109, 112, 113, 133, 183
- [125] Xiaoyan Li et al. Exploring deep anomaly detection methods based on capsule

- net. In Canadian Conference on Artificial Intelligence, pages 375–387. Springer, 2020. [85](#), [179](#), [180](#)
- [126] Zeju Li, Konstantinos Kamnitsas, and Ben Glocker. Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation. In MICCAI, pages 402–410. Springer, 2019. [2](#), [5](#), [142](#), [143](#)
- [127] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690, 2017. [14](#), [156](#)
- [128] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. IEEE transactions on pattern analysis and machine intelligence, 2018. [2](#), [5](#), [142](#), [143](#), [144](#), [148](#)
- [129] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014. [160](#)
- [130] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2152–2161, 2019. [5](#), [14](#), [154](#), [156](#), [162](#), [164](#)
- [131] Geert Litjens et al. A survey on deep learning in medical image analysis. Medical image analysis, 42:60–88, 2017. [34](#), [48](#), [96](#), [141](#)
- [132] C. Liu, X. Xu, and Y. Zhang. Temporal attention network for action proposal. In 2018 25th IEEE International Conference on Image Processing (ICIP), pages 2281–2285, 2018. [114](#)
- [133] Fengbei Liu et al. Self-supervised depth estimation to regularise semantic segmentation in knee arthroscopy. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 594–603. Springer, 2020. [4](#), [11](#), [12](#), [34](#), [48](#), [50](#), [170](#)
- [134] Fengbei Liu, Yu Tian, Filipe R Cordeiro, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Noisy label learning for large-scale medical image classification. arXiv preprint arXiv:2103.04053, 2021. [34](#), [48](#), [96](#)
- [135] Fengbei Liu, Yu Tian, et al. Self-supervised mean teacher for semi-supervised chest x-ray classification. arXiv preprint arXiv:2103.03629, 2021. [34](#), [48](#), [96](#)
- [136] Wenqian Liu et al. Towards visually explaining variational autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8642–8651, 2020. [88](#)
- [137] Wen Liu, Weixin Luo, Zhengxin Li, Peilin Zhao, Shenghua Gao, et al. Margin learning embedded prediction for video anomaly detection with a few anomalies. In IJCAI, pages 3023–3030, 2019. [13](#), [110](#), [111](#), [184](#)
- [138] Weitang Liu, Xiaoyun Wang, John D Owens, and Yixuan Li. Energy-based out-of-distribution detection. arXiv preprint arXiv:2010.03759, 2020. [154](#), [157](#), [159](#),

- 163, 164, 165, 190
- [139] Y. Liu, Y. Tian, G. Maicas, L. Z. Cheng Tao Pu, R. Singh, J. W. Verjans, and G. Carneiro. Photoshopping colonoscopy video frames. In *ISBI*, pages 1–5, 2020. 2, 4, 15, 16, 34, 35, 38, 39, 48, 49, 50, 57, 59, 88, 89, 96, 130, 142, 143, 147, 148
- [140] Y. Liu, Y. Tian, G. Maicas, L. Z. Cheng Tao Pu, R. Singh, J. W. Verjans, and G. Carneiro. Photoshopping colonoscopy video frames. In *ISBI*, pages 1–5, 2020. 34, 42, 48, 50, 59, 61, 100, 101
- [141] Yuyuan Liu, Yu Tian, Gabriel Maicas, Leonardo Zorron Cheng Tao Pu, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Photoshopping colonoscopy video frames. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2020. 110
- [142] Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. *Advances in Neural Information Processing Systems*, 32:10623–10633, 2019. 154, 159, 163, 164, 165
- [143] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 120
- [144] Weixin Luo, Zaiwang Gu, Jiang Liu, and Shenghua Gao. Encoding structure-texture relation with p-net for anomaly detection in retinal images. *ECCV*, 2020,. 34
- [145] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017. 3, 108, 109, 111, 118, 119, 184
- [146] Cheng Tao Pu LZ et al. Computer-aided diagnosis for characterisation of colorectal lesions: a comprehensive software including serrated lesions. *Gastrointestinal Endoscopy*, 2020. 34, 48, 96
- [147] Yiting Ma, Xuejin Chen, Kai Cheng, Yang Li, and Bin Sun. Ldpolypvideo benchmark: A large-scale colonoscopy video dataset of diverse polyps. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 387–396. Springer, 2021. 17, 135
- [148] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015. 4, 23, 142, 143
- [149] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *arXiv preprint arXiv:1802.10501*, 2018. 164
- [150] Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. Graph embedded pose clustering for anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 11, 110

- [151] Pedro Henrique Martins, Zita Marinho, and André FT Martins. infinity-former: Infinite memory transformer. [arXiv preprint arXiv:2109.00301](#), 2021. 97
- [152] Gérard Medioni, Isaac Cohen, François Brémond, Somboon Hongeng, and Ramakant Nevatia. Event detection and analysis from video streams. *IEEE Transactions on pattern analysis and machine intelligence*, 23(8):873–889, 2001. 110
- [153] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 11, 110
- [154] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11996–12004, 2019. 110
- [155] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. [arXiv preprint arXiv:1811.12709](#), 2018. 5, 14, 154, 156, 164
- [156] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Buló, and Peter Kotschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 189
- [157] Duc Tam Nguyen, Zhongyu Lou, Michael Klar, and Thomas Brox. Anomaly detection with multiple-hypotheses predictions. In *International Conference on Machine Learning*, pages 4800–4809. PMLR, 2019. 75, 110
- [158] Loc Nguyen. Tutorial on em algorithm. [arXiv preprint](#), 2020. 81, 177
- [159] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 110
- [160] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. [arXiv preprint arXiv:1803.02999](#), 2018. 143
- [161] Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. [arXiv preprint arXiv:1904.09770](#), 2019. 157
- [162] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 36, 37
- [163] Khalil Ouardini et al. Towards practical unsupervised anomaly detection on retinal images. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pages 225–234. Springer, 2019. 34
- [164] Guansong Pang, Longbing Cao, Ling Chen, and Huan Liu. Learning represen-

- tations of ultrahigh-dimensional data for random distance-based outlier detection. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 2041–2050, 2018. [13](#), [110](#), [144](#)
- [165] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. ACM Computing Surveys (CSUR), 54(2):1–38, 2021. [4](#), [34](#), [50](#), [76](#), [110](#)
- [166] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection: A review. arXiv preprint arXiv:2007.02500, 2020. [11](#)
- [167] Guansong Pang, Chunhua Shen, and Anton van den Hengel. Deep anomaly detection with deviation networks. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 353–362, 2019. [xix](#), [48](#), [50](#), [96](#), [110](#), [130](#), [142](#), [143](#), [145](#), [146](#)
- [168] Guansong Pang, Anton van den Hengel, Chunhua Shen, and Longbing Cao. Toward deep supervised anomaly detection: Reinforcement learning from partially labeled anomaly data. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pages 1298–1308, 2021. [130](#), [136](#)
- [169] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12173–12182, 2020. [11](#), [110](#)
- [170] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020. [12](#), [110](#), [118](#), [119](#)
- [171] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. arXiv preprint, 2017. [27](#), [147](#)
- [172] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019. [101](#), [118](#), [136](#)
- [173] Deepak Pathak, Abhijit Sharang, and Amitabha Mukerjee. Anomaly localization in topic-based analysis of surveillance videos. In 2015 IEEE Winter Conference on Applications of Computer Vision, pages 389–395, 2015. [12](#), [76](#)
- [174] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In CVPR, pages 2898–2906, 2019. [11](#), [23](#), [24](#), [26](#), [28](#), [41](#), [42](#), [59](#), [61](#), [89](#), [100](#), [101](#), [143](#), [144](#),

- 147, 148
- [175] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 12, 110
 - [176] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2898–2906, 2019. 15, 75, 85, 179, 180
 - [177] Pramuditha Perera and Vishal M Patel. Learning deep features for one-class classification. IEEE Transactions on Image Processing, 28(11):5450–5463, 2019. 74
 - [178] Hughes Perreault, Guillaume-Alexandre Bilodeau, Nicolas Saunier, and Maguelonne Héritier. Spotnet: Self-attention multi-task network for object detection. In 2020 17th Conference on Computer and Robot Vision (CRV), pages 230–237. IEEE, 2020. 115
 - [179] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: detecting small road hazards for self-driving vehicles. In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 1099–1106. IEEE, 2016. 156, 161
 - [180] L Pu, Zorron Cheng Tao, et al. Prospective study assessing a comprehensive computer-aided diagnosis for characterization of colorectal lesions: results from different centers and imaging technologies. In Journal of Gastroenterology and Hepatology, volume 34, pages 25–26. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2019. 34, 48, 143
 - [181] Leonardo Zorron Cheng Tao Pu et al. Computer-aided diagnosis for characterization of colorectal lesions: a comprehensive software including serrated lesions. Gastrointestinal Endoscopy, 2020. 130, 143
 - [182] Bharathkumar Ramachandra, Michael Jones, and Ranga Raju Vatsavai. A survey of single-scene video anomaly detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. 118
 - [183] Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, Enver Sangineto, and Nicu Sebe. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1689–1698. IEEE, 2018. 3, 108
 - [184] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In 2017 IEEE International Conference on Image Processing (ICIP), pages 1577–1581. IEEE, 2017. 3, 108
 - [185] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In Proceedings of

- the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2806–2814, 2021. [11](#), [59](#), [60](#), [61](#), [62](#), [91](#), [96](#), [100](#), [101](#)
- [186] Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. arXiv preprint arXiv:2106.03844, 2021. [48](#), [91](#)
- [187] Huamin Ren, Weifeng Liu, Søren Ingvor Olsen, Sergio Escalera, and Thomas B Moeslund. Unsupervised behavior-specific dictionary learning for abnormal event detection. In BMVC, pages 28–1, 2015. [110](#)
- [188] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. arXiv preprint arXiv:1803.09050, 2018. [144](#), [148](#)
- [189] Olaf Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015. [87](#), [176](#)
- [190] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015. [41](#), [42](#)
- [191] Lukas Ruff et al. Deep one-class classification. In International conference on machine learning, pages 4393–4402, 2018. [xvii](#), [11](#), [15](#), [74](#), [75](#), [77](#), [84](#), [85](#), [88](#), [175](#), [179](#), [180](#)
- [192] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In International conference on machine learning, pages 4393–4402, 2018. [50](#), [110](#)
- [193] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. arXiv preprint arXiv:1906.02694, 2019. [13](#), [74](#), [110](#)
- [194] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Transactions on Image Processing, 26(4):1992–2004, 2017. [75](#), [110](#)
- [195] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3379–3388, 2018. [11](#), [75](#), [87](#), [88](#), [176](#)
- [196] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018. [110](#)
- [197] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid R. Rabiee. Multiresolution knowledge distillation for

- anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14902–14912, June 2021. [75](#)
- [198] Thomas Schlegl et al. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. Medical image analysis, 54:30–44, 2019. [34](#), [35](#), [37](#), [38](#), [39](#), [42](#)
- [199] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. Medical image analysis, 54:30–44, 2019. [4](#), [11](#), [12](#), [23](#), [24](#), [26](#), [28](#), [48](#), [50](#), [59](#), [60](#), [61](#), [62](#), [74](#), [88](#), [89](#), [96](#), [100](#), [101](#), [143](#), [144](#), [147](#), [148](#)
- [200] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In International conference on information processing in medical imaging, pages 146–157. Springer, 2017. [74](#), [85](#), [87](#), [88](#), [176](#), [179](#), [180](#)
- [201] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. Neural computation, 13(7):1443–1471, 2001. [85](#), [179](#), [180](#)
- [202] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2002. [1](#)
- [203] Bernhard Schölkopf, Robert C Williamson, Alex J Smola, John Shawe-Taylor, and John C Platt. Support vector method for novelty detection. In Advances in neural information processing systems, pages 582–588, 2000. [120](#)
- [204] Philipp Seeböck, José Ignacio Orlando, Thomas Schlegl, Sebastian M Waldstein, Hrvoje Bogunović, Sophie Klimscha, Georg Langs, and Ursula Schmidt-Erfurth. Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct. IEEE transactions on medical imaging, 39(1):87–98, 2019. [34](#), [48](#), [50](#), [96](#)
- [205] Sorina Smeureanu, Radu Tudor Ionescu, Marius Popescu, and Bogdan Alexe. Deep appearance features for abnormal behavior detection in video. In International Conference on Image Analysis and Processing, pages 779–789. Springer, 2017. [11](#), [110](#)
- [206] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. arXiv, 2017. [89](#)
- [207] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pages 1857–1865, 2016. [35](#)
- [208] Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minh Jin, and Tomas Pfister. Learning and evaluating representations for deep one-class classification. arXiv preprint arXiv:2011.02578, 2020. [xvi](#), [34](#), [36](#), [37](#), [39](#), [48](#), [52](#), [57](#), [66](#), [68](#), [69](#), [91](#), [96](#)

- [209] Fahad Sohrab, Jenni Raitoharju, Moncef Gabbouj, and Alexandros Iosifidis. Sub-space support vector data description. In 2018 24th International Conference on Pattern Recognition (ICPR), pages 722–727. IEEE, 2018. [119](#), [120](#)
- [210] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. arXiv, 2014. [89](#)
- [211] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6479–6488, 2018. [xi](#), [3](#), [13](#), [16](#), [108](#), [109](#), [110](#), [111](#), [117](#), [118](#), [119](#), [120](#), [121](#), [122](#), [123](#), [130](#), [132](#), [133](#), [135](#), [136](#), [162](#)
- [212] Li Sun, Yanjun Chen, Wu Luo, Haiyan Wu, and Chongyang Zhang. Discriminative clip mining for video anomaly detection. In 2020 IEEE International Conference on Image Processing (ICIP), pages 2121–2125. IEEE, 2020. [110](#)
- [213] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1199–1208, 2018. [143](#)
- [214] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. arXiv preprint arXiv:2007.08176, 2020. [48](#), [50](#), [52](#), [54](#), [68](#), [69](#), [91](#)
- [215] Ta-Wei Tang et al. Anomaly detection neural network with dual auto-encoders gan and its industrial inspection applications. Sensors, 20(12):3336, 2020. [87](#), [176](#)
- [216] Yu Tian et al. Few-shot anomaly detection for polyp frames from colonoscopy. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 274–284. Springer, 2020. [4](#), [13](#), [34](#), [48](#), [50](#), [75](#), [96](#), [110](#), [130](#)
- [217] Yu Tian et al. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. arXiv preprint arXiv:2101.10030, 2021. [3](#), [34](#), [48](#), [50](#), [75](#), [130](#), [133](#), [135](#), [136](#)
- [218] Yu Tian, Fengbei Liu, Guansong Pang, Yuanhong Chen, Yuyuan Liu, Johan W. Verjans, Rajvinder Singh, and Gustavo Carneiro. Self-supervised multi-class pre-training for unsupervised anomaly detection and segmentation in medical images, 2021. [2](#), [75](#), [84](#)
- [219] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. arXiv preprint arXiv:2111.12264, 2021. [2](#), [12](#), [76](#), [96](#)
- [220] Yu Tian and others. Detecting, localising and classifying polyps from colonoscopy videos using deep learning. arXiv preprint arXiv:2101.03285, 2021. [34](#), [48](#)
- [221] Yu Tian, Guansong Pang, Fengbei Liu, Seon Ho Shin, Johan W Verjans, Rajvinder Singh, Gustavo Carneiro, et al. Constrained contrastive distribution learning

- for unsupervised anomaly detection and localisation in medical images. arXiv preprint arXiv:2103.03423, 2021. [xi](#), [xvi](#), [2](#), [48](#), [49](#), [50](#), [51](#), [52](#), [54](#), [59](#), [60](#), [63](#), [66](#), [67](#), [68](#), [69](#), [75](#), [84](#), [96](#), [100](#), [101](#), [102](#), [103](#), [104](#), [110](#), [130](#), [132](#)
- [222] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 4489–4497, 2015. [111](#)
- [223] Hristina Uzunova, Sandra Schultz, Heinz Handels, and Jan Ehrhardt. Unsupervised pathology detection in medical images using conditional variational autoencoders. International journal of computer assisted radiology and surgery, 14(3):451–461, 2019. [34](#)
- [224] Aaron Van den Oord et al. Conditional image generation with pixelcnn decoders. In Advances in neural information processing systems, pages 4790–4798, 2016. [85](#), [179](#), [180](#)
- [225] Jeroen C Van Rijn, Johannes B Reitsma, Jaap Stoker, Patrick M Bossuyt, Sander J Van Deventer, and Evelien Dekker. Polyp miss rate determined by tandem colonoscopy: a systematic review. American Journal of Gastroenterology, 101(2):343–350, 2006. [22](#), [143](#)
- [226] Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Revisiting multi-task learning in the deep learning era. arXiv preprint arXiv:2004.13379, 2, 2020. [14](#), [15](#), [157](#)
- [227] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017. [99](#)
- [228] Shashanka Venkataramanan, Kuan-Chuan Peng, Rajat Vikram Singh, and Abhijit Mahalanobis. Attention guided anomaly detection and localization in images. arXiv preprint arXiv:1911.08616, 2019. [12](#), [15](#), [74](#), [75](#), [76](#), [84](#), [85](#), [86](#), [87](#), [88](#), [89](#), [110](#), [176](#), [178](#), [179](#), [180](#)
- [229] Shashanka Venkataramanan, Kuan-Chuan Peng, Rajat Vikram Singh, and Abhijit Mahalanobis. Attention guided anomaly localization in images. In ECCV, pages 485–503. Springer, 2020. [4](#), [34](#), [41](#), [42](#), [48](#), [50](#), [96](#), [101](#), [102](#), [103](#), [104](#)
- [230] Tomas Vojir, Tomáš Šipka, Rahaf Aljundi, Nikolay Chumerin, Daniel Olmeda Reino, and Jiri Matas. Road anomaly detection by partial image reconstruction with segmentation coupling. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15651–15660, 2021. [14](#), [156](#)
- [231] B. Wan, Y. Fang, X. Xia, and J. Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. In 2020 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6, 2020. [13](#), [111](#), [118](#), [119](#), [136](#)
- [232] B. Wan, Y. Fang, X. Xia, and J. Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. In 2020 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6, 2020. [110](#), [117](#)

- [233] Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 8201–8211, 2019. [110](#), [119](#), [120](#)
- [234] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1386–1393, 2014. [110](#)
- [235] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Scientific Reports, 10(1):1–12, 2020. [15](#), [16](#), [49](#), [57](#), [62](#), [97](#), [101](#)
- [236] Peng Wang, Lingqiao Liu, Chunhua Shen, Zi Huang, Anton van den Hengel, and Heng Tao Shen. What’s wrong with that object? identifying images of unusual objects by modelling the detection score distribution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1573–1581, 2016. [75](#)
- [237] Shenzhi Wang, Liwei Wu, Lei Cui, and Yujun Shen. Glancing at the patch: Anomaly localization with global and local feature comparison. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 254–263, June 2021. [75](#)
- [238] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In ICML, pages 9929–9939. PMLR, 2020. [12](#), [34](#), [35](#), [50](#)
- [239] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7794–7803, 2018. [110](#), [111](#), [114](#), [115](#), [184](#)
- [240] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004. [83](#), [175](#)
- [241] Zhou Wang and et al. Multiscale structural similarity for image quality assessment. In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, volume 2, pages 1398–1402. Ieee, 2003. [35](#), [38](#), [56](#), [100](#)
- [242] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, volume 2, pages 1398–1402. Ieee, 2003. [80](#), [83](#), [175](#)
- [243] Lior Wolf et al. Unsupervised learning of the set of local maxima. arXiv preprint arXiv:2001.05026, 2020. [85](#), [179](#), [180](#)
- [244] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. arXiv preprint arXiv:2103.15808, 2021. [131](#), [132](#)
- [245] Peng Wu, jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei

- Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In European Conference on Computer Vision (ECCV), 2020. [13](#), [16](#), [108](#), [109](#), [110](#), [111](#), [115](#), [117](#), [118](#), [119](#), [120](#), [130](#), [132](#), [136](#), [184](#)
- [246] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan L Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In European Conference on Computer Vision, pages 145–161. Springer, 2020. [5](#), [14](#), [154](#), [156](#), [165](#)
- [247] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv, 2017. [15](#), [84](#)
- [248] Liang Xiong, Barnabás Póczos, and Jeff Schneider. Group anomaly detection using flexible genre models. Advances in neural information processing systems, 24, 2011. [1](#)
- [249] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. arXiv preprint arXiv:1510.01553, 2015. [110](#)
- [250] Dan Xu, Rui Song, Xinyu Wu, Nannan Li, Wei Feng, and Huihuan Qian. Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts. Neurocomputing, 143:144–152, 2014. [16](#), [111](#), [117](#), [184](#)
- [251] Xudong Yan et al. Learning semantic context from normal samples for unsupervised anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence, 2021. [85](#), [87](#), [88](#), [176](#)
- [252] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In ACCV, 2020. [36](#)
- [253] Chong You, Daniel P Robinson, and René Vidal. Provable self-representation based outlier detection in a union of subspaces. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3395–3404, 2017. [11](#)
- [254] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122, 2015. [110](#), [111](#), [114](#), [184](#)
- [255] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. arXiv preprint arXiv:2008.11988, 2020. [118](#), [119](#)
- [256] Sangdoon Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6023–6032, 2019. [55](#), [160](#)
- [257] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, Arif Mahmood, and Seung-Ik Lee. Cleaning label noise with clusters for minimally supervised anomaly detection. arXiv preprint arXiv:2104.14770, 2021. [13](#), [110](#)
- [258] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression

- for anomalous event detection. In European Conference on Computer Vision, pages 358–376. Springer, 2020. [13](#), [110](#), [136](#)
- [259] Muhammad Zaigham Zaheer, Arif Mahmood, Hochul Shin, and Seung-Ik Lee. A self-reasoning framework for anomaly detection using video-level labels. IEEE Signal Processing Letters, 27:1705–1709, 2020. [13](#), [110](#)
- [260] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem - a discriminatively trained reconstruction embedding for surface anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 8330–8339, October 2021. [75](#)
- [261] Hongyi Zhang et al. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017. [79](#)
- [262] J. Zhang, L. Qing, and J. Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In 2019 IEEE International Conference on Image Processing (ICIP), pages 4030–4034, 2019. [13](#), [109](#), [110](#), [111](#), [117](#), [118](#), [119](#), [120](#)
- [263] Tianzhu Zhang, Hanqing Lu, and Stan Z Li. Learning semantic scene models by object classification and trajectory clustering. In 2009 IEEE conference on computer vision and pattern recognition, pages 1940–1947. IEEE, 2009. [110](#)
- [264] Ying Zhang, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Shun Sakai. Video anomaly detection based on locality sensitive hashing filters. Pattern Recognition, 59:302–311, 2016. [3](#), [108](#)
- [265] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10076–10085, 2020. [11](#), [110](#), [115](#)
- [266] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1237–1246, 2019. [13](#), [16](#), [17](#), [108](#), [109](#), [111](#), [115](#), [117](#), [118](#), [119](#), [120](#), [121](#), [130](#), [136](#), [184](#)
- [267] Bolei Zhou et al. Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2921–2929, 2016. [89](#)
- [268] Kang Zhou, Yuting Xiao, Jianlong Yang, Jun Cheng, Wen Liu, Weixin Luo, Zaiwang Gu, Jiang Liu, and Shenghua Gao. Encoding structure-texture relation with p-net for anomaly detection in retinal images. In European Conference on Computer Vision, pages 360–377. Springer, 2020. [110](#)
- [269] Zongwei Zhou et al. Unet++: A nested u-net architecture for medical image segmentation. In Deep learning in medical image analysis and multimodal learning for clinical decision support, pages 3–11. Springer, 2018. [41](#), [42](#), [101](#)
- [270] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. arXiv preprint arXiv:1907.10211, 2019. [13](#), [109](#), [111](#), [119](#), [120](#)

- [271] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8856–8865, 2019. [162](#), [189](#)
- [272] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining: The ASA Data Science Journal, 5(5):363–387, 2012. [1](#)
- [273] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In International conference on learning representations, 2018. [4](#), [23](#), [110](#), [142](#), [143](#)