THE UNIVERSITY
of ADELAIDE

Weakly-supervised learning in Computer Vision and Medical Imaging

by

Fengbei Liu

A thesis submitted for the degree of

Doctor of Philosophy

February 8, 2024

Australian Institute for Machine Learning (AIML)

The University of Adelaide

Contents

List of Tables

List of Figures

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.

Fengbei Liu February 8, 2024

Acknowledgements

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

First, I would like to thank my principle supervisor, Professor Gustavo Carneiro, for his tremendous help during my Ph.D years. He has been supportive since the day we met during my honors years. His mentorship encourages me to explore new idea and consoled me during challenging times. Without his guidance, I would not have been able to succeed in my Ph.D. journey. His work ethic, research attitude, and kindness have deeply inspired me. It is a great honor to be his student.

I would also like to extend my thanks to other senior members who have assisted me during these years: Dr. Filipe Cordeiro, Professor Vasileios Belagiannis, Professor Ian Reid, and Professor Mark Jenkinson. Their support, advice, and feedback have greatly enhanced my research. I am grateful for their assistance with programming tasks, paper revisions, academic issue explanations, and research proposal finalization. Their help was invaluable during hard times.

I am also grateful for my co-authors and friends for their close discussions and collaboration. These constructive discussions have enlightened me with novel ideas and resulted in successful research projects and papers. They are Yu Tian, Yuyuan Liu, Yuanhong Chen, Chong Wang, Arpit Garg. I wish them all successful careers in the future.

Last but not least, I would like to thank my family for their unwavering support during my study. They have cherished with me every great moment and supported me whenever I need it. I will always be grateful for their accompany.

Publications

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

- **Fengbei Liu**, Yuanhong Chen, Chong Wang, Yuyuan Liu, Gustavo Carneiro, Generative Noisy-Label Learning by Implicit Dicriminative Approximation with Partial Label Prior. *Arxiv Preprint, Under Review*, 2023.

- **Fengbei Liu**, Yuanhong Chen, Chong Wang, Yu Tian, Gustavo Carneiro. Asymmetric Co-teaching with Multi-view Consensus for Noisy Label Learning. *Arxiv Preprint, Under Review*, 2023.

- Yuanhong Chen*, **Fengbei Liu***, Hu Wang, Yu Tian, Chong Wang, Yuyuan Liu, Gustavo Carneiro. BoMD: Bag of Multi-label Local Descriptors for Noisy Chest X-ray Classification. *International Conference on Computer Vision (**ICCV**),* 2023.

- **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**.

- **Fengbei Liu***, Yu Tian*, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, Gustavo Carneiro. ACPL: Anti-curriculum Pseudo Labelling for Semi-supervised Medical Image Classification *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.

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

- Chong Wang, Yuyuan Liu, Yuanhong Chen, **Fengbei Liu**, Yu Tian, Davis J McCarthy, Helen Frazer, Gustavo Carneiro. Learning Support and Trivial Prototypes

for Interpretable Image Classification. *International Conference on Computer Vision (**ICCV**),* 2023.

- 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, Hu Wang, Chong Wang, Yu Tian, **Fengbei Liu**, Yuyuan Liu, Michael Elliott, Davis J 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**.

- Chong Wang, Yuanhong Chen, Yuyuan Liu, Yu Tian, **Fengbei Liu**, Davis J 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**.

- Yuyuan Liu, Yu Tian, Yuanhong Chen, **Fengbei Liu**, Vasileios Belagiannis, Gustavo Carneiro. Perturbed and Strict Mean Teachers for Semi-supervised Semantic Segmentation. *Computer Vision and Pattern Recognition (**CVPR**),* 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. *European Conference on Computer Vision (**ECCV**),* 2022, **Oral**.

- Yu Tian*, **Fengbei Liu***, Guansong Pang, Yuanhong Chen, Yuyuan Liu, Johan W Verjans, Rajvinder Singh, Gustavo Carneiro. Self-supervised Multi-class Pretraining for Unsupervised Anomaly Detection and Segmentation in Medical Images. *Medical Image Analysis (**MIA**),* 2023.

- 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. *International Conference on Medical Imaging Computing and Computer-Assisted Intervention (**MICCAI**)*, 2021.

- Yaqub Jonmohamadi, Shahnewaz Ali, **Fengbei Liu**, Jonathan Roberts, Ross Crawford, Gustavo Carneiro, Ajay K Pandey. 3D Semantic Mapping from Arthroscopy using Out-of-distribution Pose and Depth and In-distribution Segmentation Training.

*International Conference on Medical Imaging Computing and Computer-Assisted Intervention (**MICCAI**)*, 2021.

- **Fengbei Liu**, Yaqub Jonmohamadi, Gabriel Maicas, Ajay K Pandey, Gustavo Carneiro. Self-supervised depth estimation to regularise semantic segmentation in knee arthroscopy. *International Conference on Medical Imaging Computing and Computer-Assisted Intervention (**MICCAI**)*, 2020.

- Yuanhong Chen, Yuyuan Liu, Hu Wang, **Fengbei Liu**, Chong Wang, Gustavo Carneiro. A Closer Look at Audio-Visual Semantic Segmentation. *Arxiv Preprint, Under Review*, 2023.

- Yuanhong Chen, Yuyuan Liu, Chong Wang, Michael Elliott, Chun Fung Kowk, Yu Tian, **Fengbei Liu**, Helen Frazer, Davis J McCarthy, Gustavo Carneiro. BRAIxDet: Learning to Detect Malignant Breast Lesion with Incomplete Annotations. *Arxiv Preprint, Under Review*, 2023.

- 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*, 2023.

- 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-attentional Masked Autoencoder. In *International Workshop on Machine Learning in Medical Imaging, **MICCAI-MLMI***, 2023.

Abstract

Weakly-supervised learning is a fundamental problem in computer vision and medical imaging, aiming to learn from imperfect supervision signals. Deep neural networks have been the dominant model behind current solutions, achieving great success in various application domains. Weakly-supervised learning can be formulated as: 1) incomplete supervision, where only a small subset of training samples are annotated, 2) inaccurate supervision, where some training samples are annotated with incorrect supervision, and 3) inexact supervision, where training samples are given ambiguous or indirect supervision signals. Despite the remarkable achievements of current approaches, there are still many challenges worth exploring to advance the field, particularly in real-world datasets containing non-ideal scenarios.

State-of-the-art incomplete supervision methods such as semi-supervised learning focus on consistency regularization and explore various data augmentation strategies. However, in real-world scenarios such as Medical Image Analysis (MIA), these methods fail with severe class imbalance. Moreover, few methods have been tested under a multi-label setup, which is common in MIA. Therefore, we argue that SSL methods in MIA need to be flexible enough to handle both multi-class and multi-label, as well as imbalanced learning. To address this problem, we propose two approaches that utilize self-supervised learning and pseudo-labelling to address the aforementioned issues and consequently improve semi-supervised learning performance on MIA tasks. For inaccurate supervision such as noisy label learning, the focus is mainly on balanced multi-class classification with sample selection methods. To solve noisy label learning in MIA, we propose to use a new regularization loss that considers both noisy labels and imbalanced learning for MIA. Furthermore, we utilize multi-modality information to better re-label multi-label images in MIA. Our results on MIA benchmarks show our state-of-the-art performance and effectiveness. We also understand noisy label learning from an inexact supervision perspective by learning from label sets instead of single label supervision for multi-class classification. We analyze the advantage of multi-label learning and partial label learning in noisy label learning and demonstrate the unique property of learning with label sets.

Chapter 1

Introduction

Image classification is a fundamental problem in computer vision and medical imaging, which aims to assign predefined classes to images. This task is essential for other computer vision tasks, such as object detection [53] and semantic segmentation [66], but it is challenging for automated systems. Traditional methods for supervised classification use a multi-stage approach, where they first extract handcrafted features from images using descriptors like Scale-Invariant Feature Transform (SIFT) [139] and then feed them to a trainable classifier, such as Support Vector Machine (SVM) [169]. However, these methods depend heavily on the prior knowledge of feature extraction, which limits their performance on different domain tasks [101].

In recent years, deep learning methods have emerged as a powerful tool for image classification. These methods utilize the backpropagation algorithm [100] to adjust model parameters, such as weights and biases, based on the gradient of an objective function. They employ multiple non-linear layers to automatically learn features and patterns from images. The Convolutional Neural Network (CNN) has become the mainstream architecture for deep learning [102]. A CNN consists of multiple layers with convolutional kernels that extract useful features from locally correlated data points. The output of these convolutional operations is processed by a non-linear activation function, which aids in learning semantic differences in images. This is followed by a down-sampling operation that ensures the input is invariant to geometrical distortions [102]. The introduction of the residual network (ResNet) [67], which addresses the issue of gradient vanishing through skip connections and enables the training of very deep CNNs, has significantly advanced the architectural design of CNNs. The winning entry of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2017, Squeeze-and-Excitation Networks (SENet) [71], achieved a Top-5 error rate of 2.251%, surpassing the human-level error rate of 5.1% [158]. This achievement demonstrates that deep learning methods can perform image classification in an end-to-end framework, eliminating the need for handcrafted features and significantly outperforming traditional methods.

However, supervised deep learning methods require a large number of well-annotated training samples to function effectively [40, 109], which are often costly and time-consuming to obtain. Some non-expert data sources, such as Amazon's Mechanical Turk and text captions of collected data, have been used to automatically obtain labels. However, these sources often produce incorrect labels [202]. Furthermore, in expert domains such as medical image analysis (MIA), the labeling budget often only supports the annotation of a small number of samples, leaving a large number of samples unlabeled. Moreover, the annotations may be inconsistent or ambiguous due to the expertise or preference of different human labelers [45]. Therefore, in real-world scenarios, we often encounter problems where the available supervision information is incorrect, incomplete, or imprecise. This has motivated the development of methods beyond the supervised learning setup that can learn under imperfect annotation conditions. These methods are referred to as *weakly-supervised learning* in this thesis.

Weakly-supervised learning has emerged as an active research area [240]. Traditional approaches employ Support Vector Machines (SVMs) [10, 138, 140, 160, 169, 231] to learn a discriminative hyperplane and use various techniques to assign pseudo labels to weakly-labeled samples. However, these methods often exhibit poor generalization performance on high-dimensional real-world images obtained from natural and medical imaging processes.With the advancement of deep learning, deep learning methods have become the mainstream approach for recent weakly-supervised research [240], leading to significant performance improvements on real-world computer vision and medical image analysis (MIA) datasets compared with traditional approaches. In this thesis, the primary focus is on deep learning-based weakly-supervised learning.

1.1 Weakly-supervised Setups

With the development of weakly-supervised learning, various forms of weak supervision can be categorized as: incomplete supervision, inaccurate supervision, and inexact supervision [240], as shown in Fig. 1.1.

Incomplete supervision. Incomplete supervision refers to the scenario where only a small subset of training samples are annotated, while the rest are unlabelled. Training models on the labelled subset alone is insufficient for good generalization. The key challenge is how to leverage the abundant unlabelled samples. Two major techniques are *semi-supervised learning* and *active learning*. Semi-supervised learning assigns pseudo labels to unlabelled samples using various techniques, such as pseudo-labelling [103, 153, 155, 177] or consistency learning on unlabelled samples [11, 12, 170], aiming to improve model performance with all available samples. Active learning queries human experts to obtain ground truth labels for selected unlabelled instances [47, 94, 166, 222], aiming to improve model performance with a fixed budget of queries.

Inaccurate supervision. Ideally, supervised learning learns a mapping function

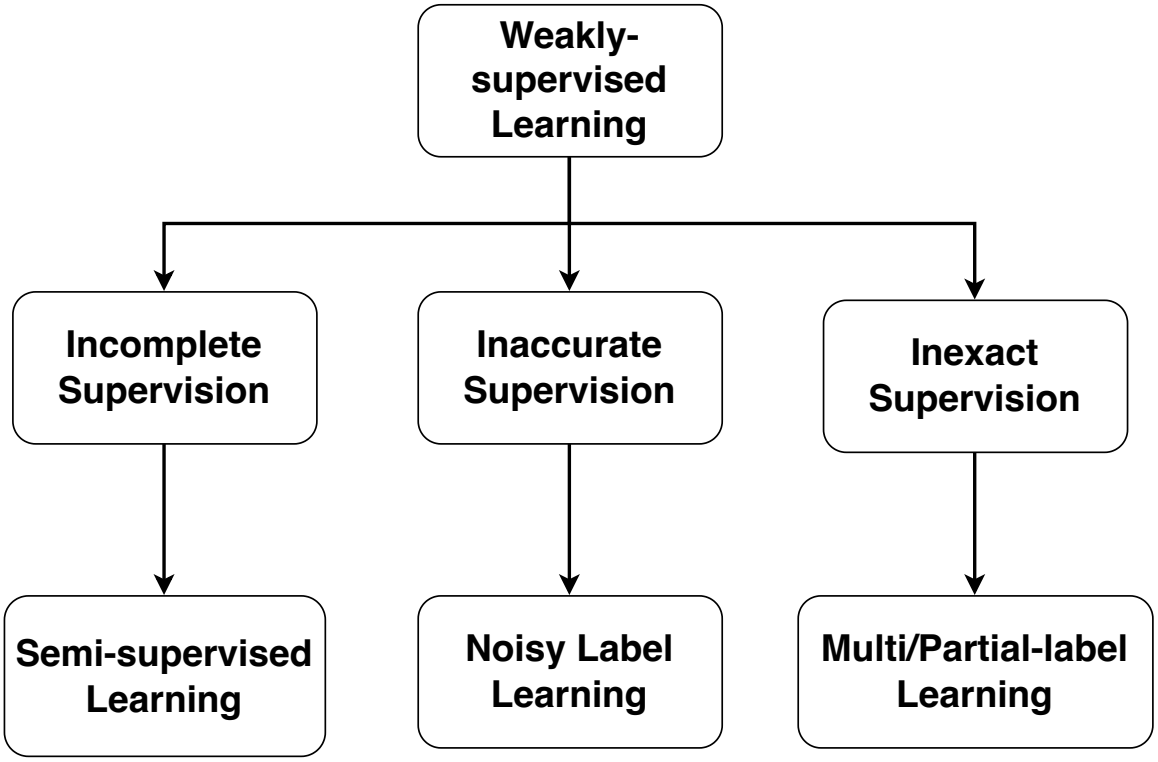


Figure 1.1: Taxonomy of different types of weakly supervision and particular methods we focused in this thesis.

between the input space and the label space. However, in practice, inaccurate labels are often encountered due to human errors or unreliable labelling sources [109, 211]. Such noisy labels can degrade the learning of the mapping function, reducing model confidence and generalization [107]. Moreover, DNNs can overfit to noisy labelled samples due to their high capacity, leading to inaccurate results on clean labelled test samples [119]. The main technique explored for this problem is *noisy-label learning*. Most noisy-label learning approaches focus on selecting clean labelled samples from a noisy training set [1, 20, 60, 81, 106, 134, 162, 198], developing novel loss functions that are robust to label noise [50, 132, 196, 235], or estimating class/instance-dependent label transition matrices [26, 148, 148, 209, 210]. The goal of noisy-label learning is to train a model under label noise and produce accurate model performance on uncorrupted test samples.

Inexact supervision. In some situations, the supervision information is not as precise as desired. A typical scenario is when each image has multiple labels, where the true label is either within or likely within the label sets. The main techniques for this problem are *partial label learning* and *multi-label learning*, depending on whether

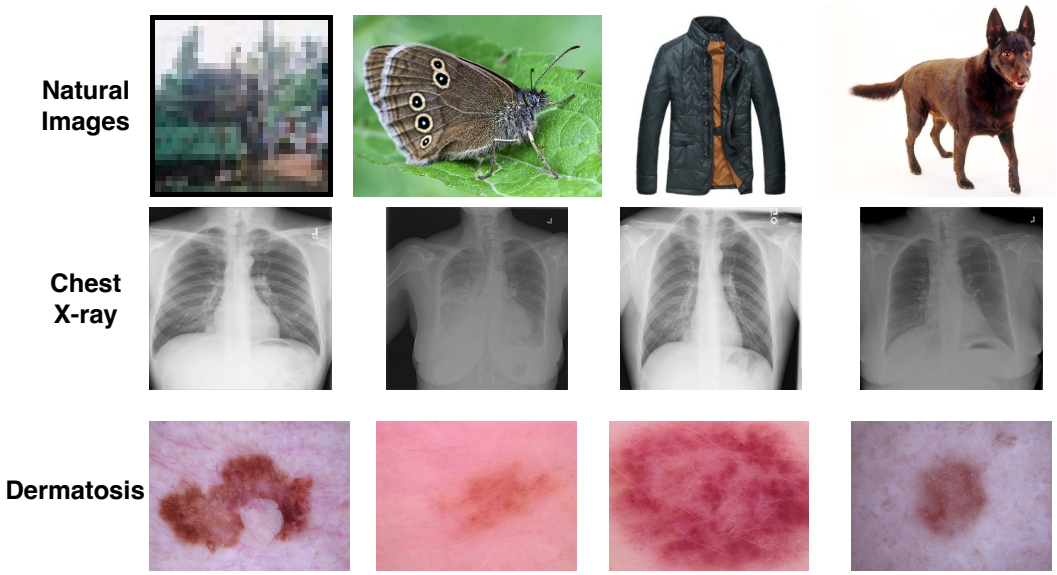


Figure 1.2: Weakly-supervised image classification problems explored in this thesis: natural images, chest x-ray images and dermatosis images.

a single-label constraint is enforced. Partial label learning extracts latent labels from label sets by either learning from all labels simultaneously [55, 56, 151, 173, 230, 239] or identifying pseudo labels for conventional multi-class classifiers [141, 186, 187, 190]. Multi-label learning removes the single-label constraint and learns each label independently [121, 154]. The goal of both tasks is to improve model performance under these inexact supervision signals.

Despite the progress of previous approaches in weakly-supervised learning, there are still significant challenges that need further investigation from both empirical and theoretical perspectives in real-world scenarios. In this thesis, we present our work under the aforementioned weakly-supervised settings: semi-supervised learning, noisy-label learning, and partial label learning. Our goal is to explore the field of saving annotation costs in dermatosis data, Chest X-ray (CXR) data, natural images data, and natural language data.

1.2 Motivation

Collecting a large number of samples with accurate and fine-grained annotations is challenging for supervised learning tasks. In real-world scenarios, weak labels are often easier to obtain. There are various forms for representing weak labels, such as a small number of labelled samples and a large number of unlabelled samples, incorrect labels extracted from search engines, or ambiguous labels with coarse supervision. There-

fore, the main motivation of weakly-supervised learning is to explore techniques for improving model performance under weak label supervision.

1.2.1 Incomplete Supervision

Semi-supervised learning (SSL) is a common approach for incomplete supervision, which deals with a large number of unlabelled samples and a small number of labelled samples. Recent SSL approaches focus on standard computer vision benchmarks with multi-class classification and class-balanced distribution. This ideal scenario enables state-of-the-art (SOTA) SSL methods to use consistency-based learning of unlabelled data [11, 12, 170] and explore strong/weak data augmentation strategies. These methods achieve outstanding performance on synthetic datasets such as CIFAR10/CIFAR100 [96]. However, in MIA real-world datasets, these methods often fail. Unlike computer vision problems, MIA has both multi-class and multi-label problems, where both have severe class imbalance issues due to the rarity of diseases. The imbalanced distribution causes pseudo labels produced by consistency learning to be biased towards majority classes and negatively affect the performance on a balanced test set. Moreover, computer vision and MIA have different domains, which makes it difficult to transfer the best augmentation strategy from computer vision to MIA. Therefore, we argue that SSL methods for real-world scenarios such as MIA need to be flexible enough to handle both multi-class and multi-label problems, as well as imbalanced learning.

To address these issues, we propose two approaches for MIA SSL. The first approach is inspired by the recent development of self-supervised learning [21, 65]. We use self-supervised pre-training in SSL to improve model performance with better representation ability. Previous MIA approaches train models from either randomly initialized weights or ImageNet [40] pre-trained weights. However, these weights are derived from natural images, which are different from MIA tasks. In our work, we use self-supervised techniques to train on massive unlabelled MIA samples and fine-tune with a small percentage of labelled samples using the mean-teacher framework [177]. Our approach with self-supervised representation and mean-teacher framework works with both multi-class (skin lesion images) and multi-label (chest x-ray images) tasks. Furthermore, the learned representation is closely related to our MIA tasks and robust under imbalanced distribution [117].

Another approach we focus on is pseudo-labeling methods. Unlike consistency-based learning, pseudo-labeling is a traditional method widely explored in semi-supervised learning [103]. It iteratively assigns high-confidence pseudo labels to unlabelled samples and trains the model jointly with labelled and pseudo-labelled samples. The model should improve its performance during this iterative training process. Pseudo-labeling methods can generalize well for both multi-class and multi-label tasks [155]. However, the main challenges for this approach are confirmation bias and threshold selection to

pseudo label an unlabelled sample. Confirmation bias causes pseudo label errors made by early model predictions to accumulate on unlabelled samples, leading to sub-optimal performance. The imbalanced distribution also affects threshold selection. Previous methods use a fixed value threshold for all classes, which encourages more pseudo label bias towards majority classes and deteriorates classification accuracy of minority classes. Inspired by the development of curriculum learning [205], we focus on selecting unlabelled samples to be pseudo-labelled instead of selecting a threshold. By comparing the distance between unlabelled and labelled samples as informativeness, we use a Gaussian Mixture Model (GMM) to automatically select unlabelled samples with the least similarity from labelled samples to be pseudo-labelled. This naturally boosts minority pseudo label selection under imbalanced distribution. Furthermore, we address confirmation bias by replacing the generation of pseudo-labels solely from the model with a more robust approach that combines a K-nearest neighbor (KNN) classifier and a conventional linear classifier for generating pseudo-labels.

1.2.2 Inaccurate Supervision

For inaccurate supervision, datasets are fully annotated but contain noisy labels. Learning with noisy labels (LNL) is a practical solution for training models on noisy training sets and maintaining performance on uncorrupted test sets. Recent LNL methods [172] use various approaches, including sample selection, noise transition matrix estimation and robust loss function design. However, these methods mostly work for balanced multi-class classification and become problematic for imbalanced multi-label classification, which is common in real-world datasets and MIA. Sample selection aims to select noisy labelled samples with early-learning phenomenon [119]. However, selecting samples with multiple annotations in LNL is not ideal since only a subset of the annotations may be wrong. Thus, sample selection in multi-label classification becomes a complicated noisy partial label learning problem that is difficult to solve. Noise transition matrix estimation provides an ideal framework for training statistically consistent classifiers [51, 210]. However, multi-label classification also needs to consider class correlation matrix, which is hard to capture without prior knowledge. Therefore, estimating noise transition matrix with class correlation matrix entangled becomes a challenging task. Robust loss function design proves that Mean Absolute Error (MAE) is a robust loss for LNL but often suffers from under-fitting issue. However, such robust loss functions only have been tested on synthetic datasets and do not achieve comparable performance on real-world datasets.

We propose two approaches for this learning with imbalanced noisy multi-label problem. Inspired by previous noisy-label learning methods and imbalanced learning [135], we propose a new robust loss function that considers both multi-label classification with imbalanced distribution and LNL. We construct a new memory module that stores non-volatile running average of model prediction logits from early-learning

stages. Furthermore, we perform logit adjustment [135] in memory storage to take into account the class prior distribution for debiasing the classification prediction estimated from the imbalanced training set. The memory module is used by a new regularization loss to penalize differences between current and early-learning model logits and regularize the training. For clean labelled samples, the new regularization loss ensures a relatively large gradient throughout the training procedure, even after fitting those samples. For noisy labelled samples, the new regularization loss reduces the gradient magnitude of supervised loss. In both scenarios, the new regularization loss removes the imbalanced class distribution impact in memory updating.

The second approach proposed for multi-label LNL is based on the exploration of multi-modality information to identify noisy labelled samples. Inspired by recent progress in multi-modal learning and large language models [41], we argue that the detection and correction of noisy multi-labelled samples can be leveraged by the semantic information present in the training labels. We propose to learn multi-label descriptors based on the projection of images into a semantic space using a set of visual descriptors. The projection process is trained by promoting the similarity between the set of visual descriptors and the semantic descriptors computed from the image's multiple labels using pre-trained language models [104]. This will help the visual descriptors become closer to clean labels in semantic space by ranking the top similarity visual descriptor of each image, we build KNN graphs for detecting noisy multi-label for training samples. The KNN graph is further used for smoothly relabel noisy training samples and re-train the model. This alleviates the aforementioned issue of sample selection in noisy multi-label learning and enables fine-grained analysis of each label in the images.

1.2.3 Noisy Label with Inexact Supervision

The process of sample selection in Label Noise Learning (LNL) primarily involves the detection and relabelling of samples labelled with noise. When a deep learning model is trained under label noise, it initially fits clean samples and subsequently overfits to noisy samples in the later stages of training. This pattern is evident in the dynamics of training loss, where the loss associated with clean samples is typically small, while that of noisy samples is significantly larger. Several sample selection methods, such as those described by Li et al. [107], leverage this phenomenon to segregate training samples into clean and noisy categories, which are then treated differently. However, these methods do not provide any assurance regarding the duration of the early-learning period. Moreover, the relabelling of noisy samples often relies on model predictions as pseudo labels, necessitating the design of complex threshold mechanisms and cross-selection processes between two independent models. We propose a simpler solution for LNL that has often been overlooked: learning from a set of labels (label-set supervision) rather than a single label. With multiple labels contained in the label-set, the model can receive supervision signals from various labels, thereby increasing the probability

of identifying latent clean labels. Furthermore, label-set supervision is less prone to confirmation bias as labels can be randomly sampled from a uniform distribution without requiring model prediction. Our argument for learning from multiple labels aligns with the concept of inexact supervision and offers a fresh perspective on this issue. The primary techniques for inexact supervision include *multi-label classification* and *partial label classification*. Unlike traditional multi-class classification, multi-label classification does not impose a single-label constraint on images and assumes that labels are independent of each other. This allows the model to learn both positive and negative labels for each image. In contrast, partial-label classification maintains a single-label constraint but only within specific label sets rather than all labels. Our focus is on understanding how these two techniques for inexact supervision can be incorporated into noisy label learning and validating our argument for learning multiple labels in multi-class LNL.

We propose two approaches augmented with inexact supervision in LNL. First, we consider multi-label learning with no single-label constraint considered. Previous methods tackle LNL on prediction disagreement [60, 224], which rely on jointly training two models to update their parameters when they disagree on the predictions of the same training samples. However, these two models are generally trained under the same strategy, which will quickly converge to select similar clean samples during training. We propose to train two models with different strategies: one trained with conventional multi-class learning with a single-label constraint and one trained with multi-label learning without a single-label constraint. Furthermore, the top-ranked prediction from multi-label learning represents potentially clean candidate labels. With predictions from multi-class classification and noisy labels, the formulation of three label views of each training sample allows us to formulate multi-view consensus for fine-grained sample selection. We show in a noisy label computer vision benchmark that our method provides substantial improvements over previous SOTA methods.

We also consider partial label learning with a single-label constraint within certain label sets under generative modeling of LNL. Previous noise transition matrix methods in LNL require estimating class/instance-dependent transition matrices. However, such estimation requires additional regularization [27] or anchor points [51]. Generative modeling provides a natural solution for regularizing the transition matrix term by optimizing the conditional generation of images $P(X|Y)$ [220]. However, due to the intractability of $P(X|Y)$, generative modeling requires a latent variable Z for controlling image generation and building $P(X|Y,Z)$. By introducing Z, an auxiliary generative module (Variational Autoencoder or Generative Adversarial model) is used, which greatly increases the training cost. Furthermore, related works [36, 159, 183] suggest that image generation does not help improve discriminative task performance. This motivates us to rethink generative modeling in LNL by restricting the power of image generation within finite training sets and bypassing the need for estimating Z. The

resulting framework allows our framework to have a similar structure to a discriminative approach but optimized for a generative goal. Furthermore, the new formulation allows us to connect with partial label learning by putting a non-uniform prior on each sample. We prove that our method achieves comparable performance in noisy label computer vision benchmarks and significantly outperforms previous generative modeling methods both in performance and efficiency.

1.3 Contributions and Thesis Outline

We propose different deep learning methods for various weakly supervised learning tasks with semi-supervised learning and noisy label learning. This thesis aims to propose new methodologies to be applied in computer vision and MIA tasks.

The contributions of this thesis can be outlined as:

- **Chapter 2** provides details of related literature. We introduce previously published weakly supervised learning methods, including consistency-based semi-supervised learning, pseudo-label-based semi-supervised learning, and self-supervised pre-training. We also review related works in noisy label learning, including sample selection-based methods, noise transition matrix methods, and robust loss functions. Furthermore, we present research fields closely related to MIA, including multi-label classification and imbalanced learning.

- **Chapter 3** describes our proposed self-supervised mean teacher framework for semi-supervised learning on medical image classification. Our framework is a two-stage framework that first pre-trains on massive unlabelled samples to obtain an effective feature representation. In the second stage, we use the mean-teacher framework and fine-tuning using self-supervised feature representation for semi-supervised model training. We show that our framework is more effective than previous consistency-based semi-supervised approaches on MIA datasets.

- **Chapter 4** focuses on revisiting the pseudo-labeling approach for the semi-supervised medical image classification task. We analyze the requirements and burdens of pseudo-label methods for medical image classification tasks. We propose to select the most informative unlabelled samples and generate pseudo labels by mixing up different classifier outputs. We show that our framework is flexible for different medical image classification tasks and that it significantly outperforms previous pseudo-labeling methods.

- **Chapter 5** presents a new regularization loss for noisy label learning with imbalanced multi-label medical image classification tasks. We analyze the gradient of clean/noisy samples during training and propose to store a non-volatile memory

of model prediction logits to regularize noisy label training. Furthermore, we adjust stored memory by calculating the class prior of imbalanced datasets. We show that our framework is capable of handling this challenging setup on multiple real-world datasets and outperforms previous noisy label approaches.

- **Chapter 6** proposes a multi-modality approach to noisy multi-label learning. We utilize semantic embedding from language models and project visual descriptors into the semantic space for clean label clustering. Furthermore, we propose to build a KNN graph on multiple visual descriptors of each image and smoothly re-label multi-label samples based on neighbouring visual descriptors. We test our method on synthetic and real-world medical datasets and results show that our ptroposed approach outperforms previous methods.

- **Chapter 7** investigates training multi-class noisy label classification without the single-label constraint. We analyze how the training of a noisy-label model with the single-label constraint in a multi-class task accelerates overfitting and propose to train each label independently as multi-label learning. Furthermore, we formulate a multi-view consensus sample selection using predictions and noisy labels, which enables fine-grained selection that traditional small-loss selection cannot achieve. We show in experiments that our new formulation outperforms previous co-teaching-based approaches in multiple computer vision benchmarks.

- **Chapter 8** argues how to formulate generative modeling efficiently in noisy label learning. We show that previous generative noisy label methods focus on improving image generation with expensive generative modules. However, image generation does not help improve discriminative task performances. We propose a simple framework with an assumption to estimate the generative term using only discriminative structures. Furthermore, our formulation allows us to place a non-uniform clean label prior for each instance derived from partial label learning. We show in multiple computer vision benchmarks that we achieve comparable performance and significantly improve upon previous generative noisy label methods both in performance and efficiency.

- **Chapter 9** summarizes this thesis contribution to weakly-supervised learning and also discuss potential new directions based on our progress.

Chapter 2

Literature Review

In this chapter, we review the weakly-supervised methods we focus on, including semi-supervised learning, noisy label learning, multi/partial label learning and imbalanced learning, together with their application on medical image analysis (MIA) tasks.

2.1 Semi-supervised Learning

Consistency regularization is a pivotal approach in semi-supervised learning, which emphasizes enforcing consistent model outputs under perturbations in the input or model space. Consistency is quantified as the discrepancy between the original model output and the perturbed output. A broad range of strategies propose generating different perturbations under input variations. Standard data augmentation is commonly employed by injecting noise or applying transformations such as cropping or flipping on image data. The Π Model [99] creates random augmentations of a sample and passes these randomly augmented samples multiple times to obtain different predictions. The consistency regularization in the Π model expects predictions from randomly augmented samples to be as consistent as possible. Temporal Ensembling [99] improves upon the Π Model by leveraging the Exponential Moving Average (EMA) of model prediction from past epochs, which reduces the computation cost for consistency regularization. Adversarial perturbation augments input samples with adversarial noise that reduces prediction confidence or changes prediction from the correct label. An example of a method that relies on adversarial perturbation is the Virtual Adversarial Training (VAT) [136] that aims to first generate an adversarial transformation of a sample which changes the model prediction. Then, consistency regularization is applied between the original model output and perturbed output. Mixup [229] performs linear interpolations of two inputs and their corresponding labels. This simple technique imposes consistency regularization to guide the learning of a mapping between the interpolated input and interpolated output to learn from unlabelled data. State-Of-The-

Art (SOTA) consistency-based regularization methods focus on learning automated augmentation strategies from data to produce relevant training samples. Such idea can be realised by enforcing consistency between a weakly-augmented sample and a strongly-augmented version of the same sample. ReMixMatch [11] introduces CTAugment [11] to learn automated augmentation policy. Unsupervised Data Augmentation (UDA) [213] adopts RandAugment [34] to sample transformation from Python Image Library. FixMatch [170] combines aforementioned CTAugment and RandAugment for boosting performance. Although consistency-based approaches achieve remarkable performance in computer vision semi-supervised learning tasks, their improvements rely on advanced domain-specific augmentation strategies and Mixup usage. For other domains such as MIA, it is still unknown which is the best strategy for data augmentation. Furthermore, for real-world datasets containing multi-labelled images, Mixup is known to be problematic [90, 189].

Pseudo-labelling is a semi-supervised learning approach that utilizes confident model predictions to generate pseudo labels for unlabelled samples. Initial attempts at pseudo-labelling encompassed Entropy Minimization (EntMin) [57] and Pseudo-label (PL) [103]. EntMin, which uses lower entropy as an indicator of higher model prediction confidence, was introduced as a regularization term to prompt the model to make low-entropy predictions for unlabelled samples. PL, on the other hand, generates pseudo labels by training the model on labelled samples and selecting the highest confidence prediction from unlabelled samples as the pseudo label. However, the use of one-hot vectors as pseudo labels could lead to the propagation of confirmation bias due to possible incorrect label assignments. To address this confirmation issue, MixMatch [12] proposed the employment of predictions across various input augmentations and the management of soft pseudo labels with a temperature hyperparameter. FixMatch [170] also addresses the same issue by assigning one-hot labels only when the model prediction's confidence scores exceed a predefined threshold. Other methods include knowledge distillation [69], where a teacher model imparts knowledge to a student model by using the soft targets from the teacher model to effectively train the student. Typically, the teacher model is either a pre-trained model or an ensemble of models. Noisy Student [214] is an iterative self-training process that initially trains the teacher model on labelled samples and assigns labels to unlabelled samples for training the student. The student is then reused as a teacher in the subsequent iteration with enhanced augmentation to improve generalization ability. Recent advancements in Uncertainty-aware Pseudo-label Selection (UPS) [155] demonstrate that the pseudo-labelling approach outperforms consistency regularization in multi-label scenarios. However, generating pseudo labels for each sample necessitates a class-wise threshold, which requires prior knowledge about the correlation, difficulties, and frequency of the dataset classes.

Self-supervised learning stands apart from consistency regularization or pseudo-labelling algorithms, as it can be trained without any label information. Initially

proposed for task-agnostic unsupervised learning, it has also been explored in the context of Semi-Supervised Learning (SSL). In SSL with self-supervised learning, all training samples, whether labelled or unlabelled, are utilized in the optimisation of an unsupervised learning problem. This is subsequently followed by supervised or semi-supervised fine-tuning with the supervision of labelled samples as a downstream task. For instance, pretext tasks create a supervision signal by classifying the geometric transformation applied to the image, such as rotations, scaling, and tiling [52, 225]. Other approaches segment each image into multiple patches and predict the order of a given cut-out patch to learn the content of images [43, 142, 197]. Recent advancements in self-supervised learning have focused on instance discrimination through contrastive learning [21, 23, 65]. The state-of-the-art in self-supervised learning optimizes the network by enforcing the positive pairs (feature embeddings of the same instance) to become closer while pushing apart the negative pairs (feature embeddings of different instances).

Semi-supervised in MIA has been extensively studied. In the consistency regularization approach, SRC-MT [118] improved upon the Mean-Teacher [177] framework by enforcing consistency on the Gram Matrix for both teacher and student models. NoTeacher [184] extended the Exponential Moving Average (EMA) process with two networks combined with a probabilistic graph model. GraphXNet [4] constructs a graph from dataset samples and assigns pseudo labels to unlabelled samples through label propagation.Compared with computer vision benchmarks, semi-supervised learning in MIA is usually multi-labelled and severely imbalanced. Therefore, it is crucial to study new approaches for handling real-world semi-supervised scenarios.

2.2 Noisy Label Learning

Sample selection is a key approach in noisy label learning, aiming to automatically classify training samples into clean or noisy categories and treat them differently during the training process. Previous studies [119, 226] have demonstrated that when training with a noisy label, Deep Neural Networks (DNNs) first fit the samples with clean labels and gradually overfit the samples with noisy labels. This characteristic of training loss has led researchers to assume that samples with clean labels have small losses, especially at early training stages – this is known as the *small-loss assumption*. For instance, M-correction [1] automatically selects clean samples by modeling the training loss distribution with a Beta Mixture Model (BMM). Sample selection has been combined with prediction disagreement in several works, such as Co-teaching [60] and Co-teaching+ [224], which train two networks simultaneously. In each mini-batch, they select small-loss samples for the training of the other model. JoCoR [198] improves upon Co-teaching+ by using a contrastive loss to jointly train both models. DivideMix [106] has advanced the field by combining sample selection and prediction disagreement us-

ing semi-supervised learning, co-teaching, and small-loss detection with a Gaussian Mixture Model (GMM). InstanceGM [48] combines a graphical model with DivideMix to achieve promising results. Other approaches select clean samples based on the K nearest neighbor classification in intermediate deep learning feature spaces [145, 195], distance to the class-specific eigenvector from the gram matrix eigen-decomposition using intermediate deep learning feature spaces [91], uncertainty measures [95], or prediction consistency between teacher and student models [87]. However, the drawback of sample selection is its lack of adaptation to real-world datasets such as multi-labelled samples or imbalanced distribution. When applying the small-loss approach to these datasets, the small loss could also represent samples from minority classes and cannot effectively select noisy samples.

Noise-robust methods utilize robust loss functions to counteract the overfitting effects induced by label noise during the training process. Early studies, such as [196], investigated the symmetric property of cross-entropy (CE) loss for noise-robust learning. Zhang et al. [235] observed that Mean-Absolute-Error (MAE) is robust to noisy labels and can be combined with conventional CE loss to strike a balance between convergence and generalization. Ma et al. [132] demonstrated that any loss function can be robust to label noise by applying a simple normalization term. Recently, [44] proposed a noise-robust Jensen-Shannon divergence (JSD) loss based on a soft transition between MAE and CE losses. Although these methods can reduce overfitting effects, they also tend to under-fit the training data. This issue has been partially addressed by the early learning regularization (ELR) [119], which proposes a regularization term that restricts the gradient from samples with corrupt labels. Despite ELR showing promising results, it faces challenges in multi-label scenarios, where different early convergence patterns of multiple labels can lead to poor performance under specific label noise conditions, as shown in the experiments.

Transition matrix methods aim to estimate the transition matrix between clean and noisy labels, to be used in the training of a clean label predictor using the noisy-label training samples. Theoretically, these methods can guarantee that the learned classifier is as optimal as training with clean labels if they meet the identifiability condition to estimate the transition matrix, which essentially states that under certain conditions about the annotators, one needs three labels per training sample [123]. Goldberger et al. [54] proposed a noise adaptation layer to estimate label transition using a few clean samples. Yao et al. [221] estimated the transition matrix using an intermediate class and a factorised matrix. Xia et al. [209] estimated a part-dependent transition matrix for complex noise conditions. Bae et al. [5] proposed a noisy prediction calibration method, which uses a transition matrix to reduce the gap between noisy prediction and clean label based on a KNN prediction. kMEDITM [220] applies manifold regularization for training an instance-dependent transition matrix. VolMin [110] regularizes the volume for the transition matrix simplex and improves the identifia-

bility of estimation. CausalNl [220] proposes a generative approach for estimating an instance-dependent transition matrix which does not require additional regularization or anchor points. All transition matrix methods presented above were designed to handle moisy multi-class problems, but for noisy multi-label learning, the transition matrix needs to be estimated together with a label correlation matrix estimation, which is a challenging task that still needs further investigation.

Noisy label in MIA has not been extensively studied. Given the high cost of acquiring and annotating large-scale MIA datasets, the field is considering more affordable automatic annotation processes by Natural Language Processing (NLP) approaches that extract multiple labels (each label representing a disease) from radiology reports [76, 192]. However, mistakes made by NLP combined with uncertain radiological findings can introduce label noise [143, 144], as can be found in NLP-annotated CXR datasets [76, 192] whose noisy multi-labels and class-imbalanced training samples can mislead supervised training processes. In this thesis, we make attempts to solve this challenging problem.

2.3 Weakly-supervised Datasets

In this thesis, we conducted experiments using several publicly available datasets. For computer vision datasets, we tested our methods on widely used benchmarks such as CIFAR10/100 [96], Red Mini-ImageNet [79], Clothing1M [211], and Animal10N [171]. We followed the same experimental protocol as a previous paper [107]. CIFAR10/100 contains 60,000 images with 10/100 classes, respectively. We followed the common split with 50,000 images for training and 10,000 for testing. We injected symmetric, asymmetric, and instance-dependent noise [209] for comparison with baselines. Symmetric noise randomly flips a given percentage of samples to a random label. Asymmetric noise defines prior knowledge about class correlation and flips a given percentage of samples to a correlated class. Instance-dependent noise [209] leverages sample-specific classification error for flipping labels into misclassified classes. We also studied CIFAR10N/CIFAR100N [202] to study real-world multi-rater noisy annotations for the original CIFAR10/100 images and we tested our framework on aggre, random1, random2, random3, worse types of noise on CIFAR10N and noisy on CIFAR100N. Red Mini-ImageNet is a real-world dataset [81] with images annotated with the Google Cloud Data Labelling Service. This dataset has 100 classes, each containing 600 images from ImageNet, where images are resized to 32×32 pixels from the original 84×84 to enable a fair comparison with other baselines [216]. Clothing1M is a real-world dataset with 100K images and 14 classes. The labels are automatically generated from surrounding text with an estimated noise ratio of 38.5%. The dataset also contains clean training, clean validation, and clean test sets with 50K, 14K, and 10K images, respectively, but we did not use the clean training and validation sets. The clean test-

ing is only used for measuring model performance. Animal10N [171] is a real-world dataset containing 10 animal species with five pairs of similar appearances (wolf and coyote, hamster and guinea pig, etc.). The training set size is 50K and testing size is 10K, where we followed the same setup as [25].

For medical datasets, we conduct experiments on NIH Chest X-ray14 [192], CheXpert [76], OpenI [37], PadChest [15], and ISIC2018 [29, 182]. NIH Chest X-ray14 contains 112,120 frontal-view Chest X-ray (CXR) images from 30,805 patients, with each image having between 0 and 14 annotated pathologies. The labels of NIH Chest X-ray14 are obtained from an NLP algorithm. The training set contains 86,524 images with a maximum of 9 labels per image. The dataset contains severe class imbalance, with the maximum/minimum images for each class being more than 50,000+ and less than 100. We follow the official train/test split and use the area under the ROC curve (AUC) as a metric. For the semi-supervised task, we report the classification result on the test set (26K samples) and different proportions of labelled samples. CheXpert has 224,316 frontal-view CXR images from 65,240 patients labelled with 14 common chest radiographic observations. The training set contains 170,958 images with a maximum of 8 labels per image. The labels are obtained from an NLP algorithm by extracting findings from radiology reports. We follow a commonly used setup and split the training set into 70% training, 20% validation, and 10% testing. OpenI dataset contains 3,999 radiology reports and 7,470 frontal/lateral-view CXR images from Indiana Network for Patient Care. We use all frontal-views images for evaluation, resulting in 3,818 images and 19 manually annotated diseases. PadChest contains 160,861 images with 27 chest radiographic observations. PadChest has a mixture of manually labelled frontal-view images (about 15.25% of the images). We use OpenI and PadChest manual label part images as a test set for noisy multi-label classification. For MIA multi-class datasets, ISIC2018 is a skin lesion dataset that contains 10,015 images with seven labels. Each image is associated with one of the labels, forming a multi-class classification problem. We follow the train/test split from SRC-MT [118] for a fair comparison, where the training set contains 20% labelled samples and 80% unlabelled samples. We report the AUC, Sensitivity, and F1 score results.

Statement of Authorship

Title of Paper	Self-supervised Mean Teach for Semi-supervised Chest X-ray Classification
Publication Status	☐ Published ☑ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Published at International Workshop on Machine Learning in Medical Imaging, MICCAI-MLMI, 2021

Principal Author

Name of Principal Author (Candidate)	Fengbei Liu
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper and revision
Overall percentage (%)	80
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 09/14/2023

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 in 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	Yu Tian
Contribution to the Paper	Conducted experiments and wrote the revision
Signature	Date 09/15/2023

Name of Co-Author	Filipe R. Cordeiro		
Contribution to the Paper	Discussion and commented on the revision		
Signature		Date	09/14/2023

Name of Co-Author	Vasileios Belagiannis		
Contribution to the Paper	Discussion and commented on the revision		
Signature		Date	19.09.2023

Name of Co-Author	Ian Reid		
Contribution to the Paper	Discussion and commented on the revision		
Signature		Date	21/9/23

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	Discussion and wrote the revision		
Signature		Date	22/09/2023

Chapter 3

Self-supervised Mean-teacher for Semi-supervised Chest X-ray Classification

Abstract

The training of deep learning models generally requires a large amount of annotated data for effective convergence and generalisation. However, obtaining high-quality annotations is a laboursome and expensive process due to the need of expert radiologists for the labelling task. The study of semi-supervised learning in medical image analysis is then of crucial importance given that it is much less expensive to obtain unlabelled images than to acquire images labelled by expert radiologists. Essentially, semi-supervised methods leverage large sets of unlabelled data to enable better training convergence and generalisation than using only the small set of labelled images. In this paper, we propose Self-supervised Mean Teacher for Semi-supervised (S^2MTS^2) learning that combines self-supervised mean-teacher pre-training with semi-supervised fine-tuning. The main innovation of S^2MTS^2 is the self-supervised mean-teacher pre-training based on the joint contrastive learning, which uses an infinite number of pairs of positive query and key features to improve the mean-teacher representation. The model is then fine-tuned using the exponential moving average teacher framework trained with semi-supervised learning. We validate S^2MTS^2 on the multi-label classification problems from Chest X-ray14 and CheXpert, and the multi-class classification from ISIC2018, where we show that it outperforms the previous SOTA semi-supervised learning methods by a large margin. Our code is available at https://github.com/FBLADL/semi-chest.

3.1 Introduction

Deep learning has shown outstanding results in medical image analysis problems [86, 105, 112, 115, 178, 180]. However, this performance usually depends on the availability of labelled datasets, which is expensive to obtain given that the labelling process requires expert doctors. This limitation motivates the study of semi-supervised learning (SSL) methods that train models with a small set of labelled data and a large set of unlabelled data.

The current state-of-the-art (SOTA) SSL is based on pseudo-labelling methods [103, 155], consistency-enforcing approaches [11, 99, 177], self-supervised and semi-supervised learning (S^4L) [23, 225], and graph-based label propagation [4]. Pseudo-labelling is an intuitive SSL technique, where confident predictions from the model are transformed into pseudo-labels for the unlabelled data, which are then used to re-train the model [103]. Consistency-enforcing regularisation is based on training for a consistent output given model [118, 177] or input data [11, 99] perturbations. S^4L methods are based on self-supervised pre-training [22, 64], followed by supervised fine-tuning using few labelled samples [23, 225]. Graph-based methods rely on label propagation on graphs [4]. Recently, Yang et al. [219] suggested that self-supervision pre-training provides better feature representations than consistency-enforcing approaches in SSL. However, previous S^4L approaches use only the labelled data in the fine-tuning stage, missing useful training information present in the unlabelled data. Furthermore, self-supervised pre-training [22, 64] tends to use limited amount of samples to represent each class, but recently, Cai et al. [17] showed that better representation can be obtained with an infinite amount of samples. Also, recent research [155] suggests that the student-teacher framework, such as the mean-teacher [177], works better in multi-label semi-supervised tasks than other SSL methods. We speculate that this is because other methods are usually designed to work with softmax activation that only works in multi-class problems, while mean-teacher [177] does not have this constraint and can work in multi-label problems.

In this paper, we propose a self-supervised mean-teacher for semi-supervised (S^2MTS^2) learning approach that combines S^4L [23, 225] with consistency-enforcing learning based on the mean-teacher algorithm [177]. The main contribution of our method is the self-supervised mean-teacher pre-training with the joint contrastive learning [17]. To the best of our knowledge, this is the first approach, in our field, to train the mean teacher model with self-supervised learning. This model is then fine-tuned with semi-supervised learning using the exponential moving average teacher framework [177]. We evaluate our proposed method on the thorax disease multi-label datasets ChestX-ray 14 [192] and CheXpert [76], and on the multi-class skin condition dataset ISIC2018 [29, 182]. We show that our method outperforms the SOTA on semi-supervised learning [4, 59, 118, 184]. Moreover, we investigate each component of our framework for their contribution to the overall model in the ablation study.

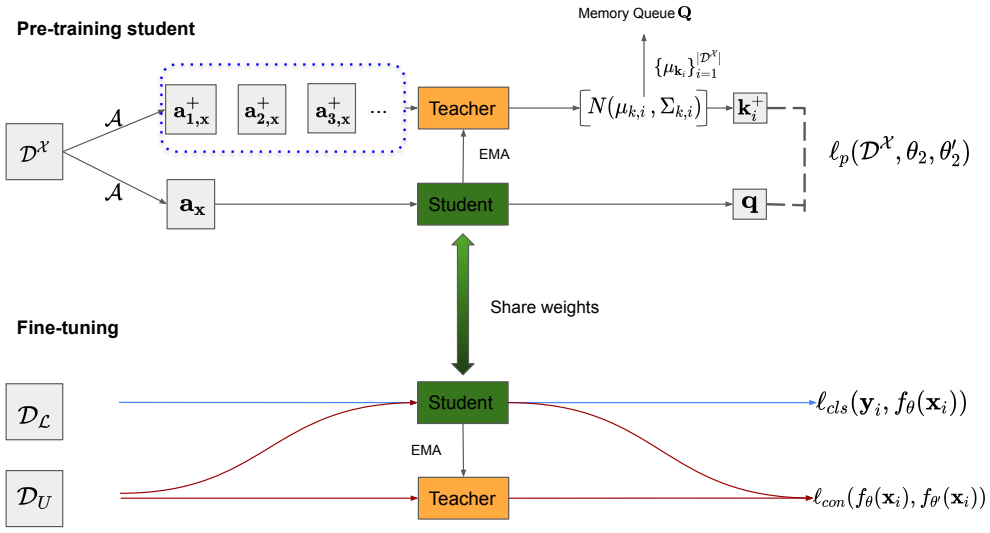


Figure 3.1: Description of the proposed self-supervised mean-teacher for semi-supervised (S^2MTS^2) learning. The main contribution of the paper resides in the top part of the figure, with the self-supervised mean-teacher pre-training based on joint contrastive learning, which uses an infinite number of pairs of positive query and key features sampled from the unlabelled images to minimise $\ell_p(.)$ in (3.1). This model is then fine-tuned with the exponential moving average teacher in a semi-supervised learning framework that uses both labelled and unlabelled sets to minimise $\ell_{cls}(.)$ and $\ell_{con}(.)$ in (3.2).

3.2 Related Works

SSL is a research topic that is gaining attention from the medical image analysis community due to the expensive image annotation process [28] and the growing number of large-scale datasets available in the field [192]. The current SOTA SSL methods are based on consistency-enforcing approaches that leverage the unlabelled data to regularise the model prediction consistency [99, 177]. Other related papers [35] extend the mean teacher [177] to encourage consistency between the prediction by the student and teacher models for atrium and brain lesion segmentation. The SOTA SSL method on Chest X-ray images [118] exploits the consistency in the relations between labelled and unlabelled data. None of these methods explores a self-supervised consistency-enforcing method to pre-train an SSL model, as we propose. Self-supervised learning methods [22, 64] are also being widely investigated in SSL because they can provide good representations [23, 225]. However, these methods ignore the large amount of unlabelled data to be used during SSL, which may lead to unsatisfactory generalisation process. An important point in self-supervised learning is on how to define the classes to be learned. In general, each class is composed of a single pair of augmented images from the same image, and many pairs of augmentations from different images [17, 22, 23, 64]. The use of a single pair of images to form a class has been criticised by Cai et al. [17], who propose the joint contrastive learning (JCL), which is an efficient way to form a class with an infinite number of augmented images from the same image to leverage the statistical dependency between different augmentations.

3.3 Method

In this section, we introduce our two-stage learning framework in detail (see Fig. 3.1). We assume that we have a small labelled dataset, denoted by $\mathcal{D}_L = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{D}_L|}$, where the image is represented by $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{H \times W \times C}$, and class $\mathbf{y} \in \{0,1\}^{|\mathcal{Y}|}$, where $\mathcal{Y}$ represents the label set. We consider a multi-label problem and thus $\sum_{c=1}^{|\mathcal{Y}|} \mathbf{y}_i(c) \in [0, |\mathcal{Y}|]$. The unlabelled dataset is defined by $\mathcal{D}_U = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}_U|}$ with $|\mathcal{D}_L| << |\mathcal{D}_U|$.

Our model consists of a student and a teacher model [177], denoted by parameters $\theta, \theta' \in \mathbf{\Theta}$, respectively, which parameterize the classifier $f_\theta : \mathcal{X} \rightarrow [0,1]^{|\mathcal{Y}|}$. This classifier can be decomposed as $f_\theta = h_{\theta_1} \circ g_{\theta_2}$, with $g_{\theta_2} : \mathcal{X} \rightarrow \mathcal{Z}$ and $h_{\theta_1} : \mathcal{Z} \rightarrow [0,1]^{|\mathcal{Y}|}$. The first stage (top of Fig. 3.1) of the training consists of a self-supervised learning that uses the images from $\mathcal{D}_L$ and $\mathcal{D}_U$, denoted by $\mathcal{D}^{\mathcal{X}} = \{\mathbf{x}_i | \mathbf{x}_i \in \mathcal{D}_L^{\mathcal{X}} \bigcup \mathcal{D}_U\}_{i=1}^{|\mathcal{D}_{\mathcal{X}}|}$, with $\mathcal{D}_L^{\mathcal{X}}$ representing the images from the set $\mathcal{D}_L$, where our method minimises the joint contrastive learning loss [17], defined in (3.1). This means that during this first stage, we only learn the parameters for g_{θ_2}. The second stage (bottom of Fig. 3.1) fine-tunes this pre-trained student-teacher model using the semi supervised consistency loss defined in (3.2). Below we provide details on the losses and training.

3.3.1 Joint Contrastive Learning to Self-supervise the Mean-teacher Pre-training

The self-supervised pre-training of the mean-teacher using joint contrastive learning (JCL) [17], presented in this section, is the main technical contribution of this paper. The teacher and student process an input image to return the keys $\mathbf{k} \in \mathcal{Z}$ and the queries $\mathbf{q} \in \mathcal{Z}$ with $\mathbf{k} = g_{\theta_2'}(\mathbf{x})$ and $\mathbf{q} = g_{\theta_2}(\mathbf{x})$. We also assume that we have a set of augmentation functions, i.e., random crop and resize, rotation and Gaussian blur, denoted by $\mathcal{A} = \{a_l : \mathcal{X} \to \mathcal{X}\}_{l=1}^{|\mathcal{A}|}$. Then JCL minimises the following loss [17]:

$$\ell_p(\mathcal{D}^{\mathcal{X}}, \theta_2, \theta_2') = -\frac{1}{|\mathcal{D}^{\mathcal{X}}|}\frac{1}{M}\sum_{i=1}^{|\mathcal{D}^{\mathcal{X}}|}\sum_{m=1}^{M}\left[\log\frac{\exp\left[\frac{1}{\tau}\mathbf{q}_i^\top\mathbf{k}_{i,m}^+\right]}{\exp\left[\frac{1}{\tau}\mathbf{q}_i^\top\mathbf{k}_{i,m}^+\right] + \sum_{j=1}^{K}\exp\left[\frac{1}{\tau}\mathbf{q}_i^\top\mathbf{k}_{i,j}^-\right]}\right], \quad (3.1)$$

where τ is the temperature hyper-parameter, the query $\mathbf{q}_i = g_{\theta_2}(a(\mathbf{x}_i))$, with $a \in \mathcal{A}$. the positive key $\mathbf{k}_{i,m}^+ \sim p(\mathbf{k}_i^+)$, with $p(\mathbf{k}_i^+) = \mathcal{N}(\mu_{\mathbf{k}_i}, \Sigma_{\mathbf{k}_i})$ and $\mathbf{k}_i = g_{\theta_2'}(a(\mathbf{x}_i))$ (i.e., a sample from the data augmentation distribution for $\mathbf{x}$), and the negative keys $\mathbf{k}_{i,j}^- \in \{\mu_{\mathbf{k}_j}\}_{i,j\in\{1,..,|\mathcal{D}^{\mathcal{X}}|\},i\neq j}$ represents a negative key for query $\mathbf{q}_i$. In (3.1), M denotes the number of positive keys, and Cai et al. [17] describe a loss that minimises a bound to (3.1) for $M \to \infty$ – below, the minimisation of $\ell_p(.)$ in (3.1) is realised by the minimisation of this bound. As defined above, the generative model $p(\mathbf{k}_i^+)$ is denoted by the Gaussian $\mathcal{N}(\mu_{\mathbf{k}_i}, \Sigma_{\mathbf{k}_i})$, where the mean $\mu_{\mathbf{k}_i}$ and covariance $\Sigma_{\mathbf{k}_i}$ are estimated from a set of keys $\{\mathbf{k}_{i,l}^+ = g_{\theta_2'}(a_l(\mathbf{x}_i))\}_{a_l\in\mathcal{A}}$ formed by different views of $\mathbf{x}_i$. The set of negative keys $\{\mu_{\mathbf{k}_j}\}_{i,j\in\{1,..,|\mathcal{D}^{\mathcal{X}}|\},i\neq j}$ is stored in a memory queue [64] that is updated in a first-in-first-out way, where the mean of the keys in $\{\mu_{\mathbf{k}_i}\}_{i=1}^{|\mathcal{D}^{\mathcal{X}}|}$ are inserted to the memory queue to replace the oldest key means from previous training iterations. The memory queue has been designed to increase the number of negative samples without sacrificing computation efficiency. The training of the student-teacher model [64, 177, 237] is achieved by updating the student parameter using the loss in (3.1), as in $\theta_2(t) = \theta_2(t-1) - \nabla_{\theta_2}\ell_p(\mathcal{D}^{\mathcal{X}}, \theta_2, \theta_2')$, where t is the training iteration. The teacher model parameter is updated with exponential moving average (EMA) with $\theta_2'(t) = \alpha\theta_2'(t-1) + (1-\alpha)\theta_2(t)$, with $\alpha \in [0, 1]$. For this pre-training stage, we notice that training for more epochs always improve the model regularisation given that it is difficult to overfit the training set with the loss in (3.1). Hence, we select the last epoch student model $g_{\theta_2}(.)$ to initialise the fine-tuning stage, defined below in Sec. 3.3.2.

3.3.2 Fine-tuning the Mean Teacher

To fine tune the mean teacher, we follow the approach in [64, 177] using the following loss to train the student model:

$$\ell_t(\mathcal{D}_L, \mathcal{D}_U, \theta, \theta') = \frac{1}{|\mathcal{D}_L|}\sum_{(\mathbf{x}_i,\mathbf{y}_i)\in\mathcal{D}_L}\ell_{cls}(\mathbf{y}_i, f_\theta(\mathbf{x}_i)) + \frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}_i\in\mathcal{D}}\ell_{con}(f_\theta(\mathbf{x}_i), f_{\theta'}(\mathbf{x}_i)), \quad (3.2)$$

where $\ell_{cls}(\mathbf{y}_i, f_\theta(\mathbf{x}_i)) = -\mathbf{y}_i^\top \log(f_\theta(\mathbf{x}_i))$, $\ell_{con}(f_\theta(\mathbf{x}_i), f_{\theta'}(\mathbf{x}_i)) = \|f_\theta(\mathbf{x}_i) - f_{\theta'}(\mathbf{x}_i)\|^2$, and $\mathcal{D} = \mathcal{D}_U \bigcup \mathcal{D}_L^{\mathcal{X}}$. The training of the student-teacher model [64, 177, 237] is achieved by updating the student parameter using the loss in (3.2), as in $\theta(t) = \theta(t-1) - \nabla_\theta \ell_t(\mathcal{D}_L, \mathcal{D}_U, \theta, \theta')$, where t is the training iteration. The teacher model parameter is updated with exponential moving average (EMA) with $\theta'(t) = \alpha\theta'(t-1) + (1-\alpha)\theta(t)$, with $\alpha \in [0, 1]$. After finishing the fine-tuning stage, we select the teacher model $f_{\theta'}(.)$ to estimate the multi-label classification for test images.

3.4 Experiment

3.4.1 Dataset Setup

We use Chest X-ray14 [192], CheXpert [76] and ISIC2018 [29, 182] datasets to evaluate our method.

Chest X-ray14 contains 112,120 chest x-ray images from 30,805 different patients. There are 14 different labels (each label represents a disease) in the dataset, where each patient can have multiple diseases at the same time, forming a multi-label classification problem. To compare with previous papers [4, 118], we adopt the official train/test data split. For the self-supervised pre-training of the mean teacher, we used all the unlabelled images (86k samples) from the training set. For the semi-supervised fine-tuning of the mean teacher, we follow the papers [4, 118] and experiment with training sets containing different proportions of labelled data (2%,5%,10%,15%,20%). We report the classification result on the official test set (26,000 samples) using area under the ROC curve (AUC).

CheXpert contains around 220,000 images with 14 different diseases, and similarly to Chest X-ray14, each patient can have multiple diseases at the same time. For pre-processing, we remove all lateral view images and treat uncertain label as negative labels. We follow the semi-supervised setup from [59], and experiment with 100/200/300/ 400/500 labelled samples per class. We report results on the official test set using AUC.

ISIC2018 is a multi-class skin condition dataset that contains 10,015 images with seven different labels. Each image is associated with one of the seven labels, forming a multi-class classification problem. We follow [118] train/test split for fair comparison, where the training contains 20% of the samples labelled, and the remaining 80% unlabelled. We report the AUC, Sensitivity, and F1 score results to compare with baselines.

3.4.2 Implementation Details

For all datasets, we use DenseNet121 [72] as our backbone model. For self-supervised pre-training, we follow [23] and replace the two-layer multi-layer perceptron (MLP) projection head by a three-layer MLP. For dataset pre-processing, we resized Chest X-ray14 images to 512 × 512 for faster processing and CheXpert and ISIC2018 to 224 × 224 for fair comparison with baselines. We use the data augmentation proposed in [22], consisting of random resize and crop, random rotation, random horizonntal flipping, except for random grayscale because X-ray images are originally in grayscale. The batch size is 128 for Chest X-ray14 and 256 for CheXpert and ISIC2018, and learning rate is 0.05. For the fine-tuning stage, we use batch size 32 with 16 labelled and 16 unlabelled. The fine-tuning takes 30 epochs with learning rate decayed by 0.1 at 15 and 25 epochs for all datasets. We use the SGD optimiser with 0.9 momentum for the pre-training stage, and Adam optimiser in fine-tuning stage. The code is written in Pytorch [147]. We use 4 Nvidia Volta-100 for the self-supervised stage and 1 Nvidia RTX 2080ti for fine-tuning.

3.4.3 Experimental Results

We evaluate our approach on the official test set of ChestX-ray14 using different percentage of labelled training data (i.e., 2%, 5%, 10%, 15%, 20%), as shown in Table 3.1. The set of labelled data used for each percentage above follows the same strategy of previous works [4, 118]. Our S^4L achieves the SOTA AUC results on all different percentages of labels. Our model surpasses the previous SOTA SRC-MT [118] by a large margin of 8.7% and 6.8% AUC for the 2% and 5% labelled set cases, respectively, where we use a backbone architecture of lower complexity (Densenet121 instead of the DenseNet169 of [118]). Using the same Densenet121 backbone, GraphXnet [4] fails to classify precisely for the 2% and 5% labelled set cases. Our method surpasses GraphXnet by more than 20% AUC in both cases. Furthermore, we achieve the SOTA results of the field for the 10%, 15% and 20% labelled set cases, outperforming all previous semi-supervised methods [4, 118]. It is worth noting that our model trained with 5% of the labelled set achieves better results than SRC-MT with 15% of labelled.We also compare with a recently proposed self-supervised pre-training methods, MoCo V2 [24], adapted to our semi-supervised task, followed by the fine-tuning stage using different percentages of labelled data. Our method outperforms MoCo V2 by almost 10% AUC when using 2% of labelled set, and almost 3% AUC for 10% of labelled set. Our result for 20% labelled set achieves comparable 81.06% AUC performance as the supervised learning approaches – 81.20% from MoCo V2 (Densenet 121) and 81.75% from SRC-MT (Densenet 169) using 100% of the labelled samples. Such result indicates the effectiveness of our proposed S^2MTS^2 in SSL benchmark problems.

We also show the class-level performance using 20% of the labelled data and com-

Label Percentage	2%	5%	10%	15%	20%	100%
Graph XNet* [4]	53.00	58.00	63.00	68.00	78.00	N/A
SRC-MT* [118]	66.95	72.29	75.28	77.76	79.23	81.75
NoTeacher [184]	72.60	77.04	77.61	N/A	79.49	N/A
MOCO V2 [24]	65.97	73.84	77.07	79.37	80.17	81.20
Ours	**74.69**	**78.96**	**79.90**	**80.31**	**81.06**	**82.51**

Table 3.1: Mean AUC result over the 14 disease classes of Chest X-Ray14 for different label set training percentages. * indicates the methods that use Densenet169 as backbone architecture.

pare with other SOTA methods in Tab. 3.2. We compare with the previous baselines, namely original mean teacher (MT) with Densenet169, SRC-MT with Densenet169, MoCo V2, and GraphXNet with Densenet121. We also train a baseline Densenet121 model with 20% labelled data using Imagenet pre-trained model. Our method achieves the best results on nine classes, surpassing the original MT [177] and its extension SRC-MT [118] by a large margin, demonstrating the effectiveness of our self-supervised learning.

Furthermore, we compare our approach on the fully-supervised Chest X-ray14 benchmark in Tab. 3.3. To the best of our knowledge, Hermoza et al. [68] has the SOTA supervised classification method containing a complex structure (relying on the weakly-supervised localisation of lesions) with a mean AUC of 82.1% (over the 14 classes), while ours reports a mean AUC of 82.5%. Hence, our model, using the whole labelled set, achieves the SOTA performance on 8 classes and an average that surpasses the previous supervised methods by a minimum of 0.4% and a maximum of 8% AUC. The results on CheXpert and ISIC2018 datasets are shown in Tables 3.4 and 3.5, respectively. In particular, for CheXpert in Table 3.4, we compare our method with LatentMixing [59] and our result is better in all cases. For ISIC2018 on Table 3.5, using the test set from SRC-MT [118], our method outperforms all baselines (Supervised, MT, and SRC-MT) for all measures.

3.4.4 Ablation Study

We study the impact of different components of our proposed S^2MTS^2 in Tab. 3.6 using Chest X-Ray14. Using the proposed self-supervised learning with just the student model, our model achieves at least 71.95% mean AUC on various percentages of labelled training data. Adding the JCL component improves the baseline by around 1% mean AUC on each training percentage. Adding the mean teacher boosts the result by 1.5% to 2% mean AUC on each training percentage. The combination of all our proposed three components achieves SOTA performance on semi-supervised task.

Method	Densenet-121	GraphXNet [4]	MOCO V2 [24]	MT [118] *	SRC-MT [118] *	Ours
Atelectasis	75.75	71.89	77.21	75.12	75.38	**78.57**
Cardiomegaly	80.71	87.99	85.84	87.37	87.7	**88.08**
Effusion	79.87	79.2	81.62	80.81	81.58	**82.87**
Infiltration	69.16	**72.05**	70.91	70.67	70.4	70.68
Mass	78.40	80.9	81.71	77.72	78.03	**82.57**
Nodule	74.49	71.13	**76.72**	73.27	73.64	76.60
Pneumonia	69.55	**76.64**	71.08	69.17	69.27	72.25
Pneumothorax	84.70	83.7	85.92	85.63	86.12	**86.55**
Consolidation	71.85	73.36	74.47	72.51	73.11	**75.47**
Edema	81.61	80.2	83.57	82.72	82.94	**84.83**
Emphysema	89.75	84.07	91.10	88.16	88.98	**91.88**
Fibrosis	79.30	80.34	80.96	78.24	79.22	**81.73**
Pleural Thicken	73.46	75.7	75.65	74.43	75.63	**76.86**
Hernia	86.05	87.22	85.62	**87.74**	87.27	85.98
Mean	78.19	78.88	80.17	78.83	79.23	**81.06**

Table 3.2: Class-level AUC comparison between our S^2MTS^2 and other semi-supervised SOTA approaches trained with **20% of labelled data** on Chest X-Ray14. * denotes the methods that use Densenet-169 as backbone.

Method	Wang et al. [192]	Li et al. [112]	CheXNet [152]	CRAL [58]	Ma et al. [131]	Hermoza et al. [68]	Ours
Atelectasis	70	72.9	75.5	78.1	77.7	77.5	**78.7**
Cardiomegaly	81	84.6	86.7	88.3	**89.4**	88.1	87.4
Effusion	75.9	78.1	81.5	83.1	82.9	83.1	**83.8**
Infiltration	66.1	67.3	69.4	69.7	69.6	69.5	**70.9**
Mass	69.3	74.3	80.2	83	**83.8**	82.6	83.3
Nodule	66.9	75.8	73.5	76.4	77.1	78.9	**79.9**
Pneumonia	65.8	63.3	69.8	72.5	72.2	**74.1**	73.9
Pneumothorax	79.9	79.3	82.8	86.6	86.2	**87.9**	87.1
Consolidation	70.3	72	72.2	75.8	75	74.7	**75.9**
Edema	80.5	71	83.5	85.3	84.6	**84.6**	84.5
Emphysema	83.3	75.1	85.6	91.1	90.8	93.6	**93.7**
Fibrosis	78.6	76.1	80.3	82.6	82.7	83.3	**83.4**
Pleural Thicken	68.4	73	74.9	78	77.9	79.3	**79.3**
Hernia	87.2	66.8	89.4	91.8	**93.4**	91.7	93.3
Mean	74.5	73.9	78.9	81.6	81.7	82.1	**82.5**

Table 3.3: Class-level AUC comparison between our S^2MTS^2 and other supervised SOTA approaches trained with **100% of labelled data** on Chest X-Ray14.

3.5 Conclusion

In this paper, we presented a novel semi-supervised framework, the Self-supervised Mean Teacher for Semi-supervised (S^2MTS^2) learning. The main contribution of S^2MTS^2 is the self-supervised mean teacher pre-trained based on joint contrastive learning [17], using an infinite number of pairs of positive query and key features. This model is then fine-tuned with the exponential moving average teacher framework. S^2MTS^2 is validated on the thorax disease multi-label classification problem from the datasets Chest X-ray14 [192] and CheXpert [76], and the multi-class classification from

Labelled	100	200	300	400	500
LatentMixing [59]	65.12	66.41	67.39	67.96	68.47
Ours	**66.15**	**67.85**	**70.83**	**71.37**	**71.58**

Table 3.4: Mean AUC result (over the 14 disease classes) on CheXpert for different number of training samples per class.

Method	AUC	Sensitivity	F1
Supervised	90.15	65.50	52.03
MT	92.96	69.75	59.10
SRC-MT [118]	93.58	71.47	60.68
Ours	**94.71**	**72.14**	**62.67**

Table 3.5: AUC, Sensitivity and F1 result on ISIC2018 using 20% of labelled training samples.

Self-supervised	JCL	MT	AUC (2%)	AUC (5%)	AUC (10%)	AUC (15%)	AUC (20%)
✓			71.95	76.82	78.54	79.28	80.14
✓	✓		72.60	77.46	79.18	79.83	80.62
✓		✓	73.80	77.66	79.08	79.70	80.57
✓	✓	✓	74.69	78.96	79.90	80.31	81.06

Table 3.6: Ablation studies of our method with different components on Chest X-Ray14. "Self-supervised" indicates the traditional self-supervised learning with contrastive loss [64]. "JCL" replaces contrastive loss with (3.1), "MT" stands for fine-tuned with student-teacher learning instead only fine-tuned on only labelled samples.

the skin condition dataset ISIC2018 [29, 182]. The experiments show that our method outperforms the previous SOTA semi-supervised learning methods by a large margin in all benchmarks containing a varying percentage of labelled data. We also show that the method holds the SOTA results on Chest X-ray14 [192] even for the fully-supervised problem. The ablation study shows the importance of three main components of the method, namely self-supervised learning, JCL, and the mean-teacher model. We will investigate the performance of our method on other semi-supervised medical imaging benchmarks in the future.

Acknowledgement This work was supported by Australian Research Council through grants DP180103232 and FT190100525.

Statement of Authorship

Title of Paper	ACPL: Anti-curriculum Pseudo Labelling for semi-supervised Medical Image Classification
Publication Status	☐ Published ☑ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Published at Computer Vision and Pattern Recognition (CVPR), 2022

Principal Author

Name of Principal Author (Candidate)	Fengbei Liu		
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper and revision		
Overall percentage (%)	80		
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 pap		
Signature		Date	09/14/2023

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 in 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	Yu Tian		
Contribution to the Paper	Conducted experiments and wrote the revision		
Signature		Date	09/15/2023

Name of Co-Author	Yuanhong Chen		
Contribution to the Paper	Discussion and commented on the revision		
Signature		Date	

Name of Co-Author	Yuyuan Liu		
Contribution to the Paper	Discussion and commented on the revision		
Signature		Date	

Name of Co-Author	Vasileios Belagiannis		
Contribution to the Paper	Discussion and commented on the revision		
Signature		Date	19.09.2023

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	Discussion and wrote the revision		
Signature		Date	22/09/20233

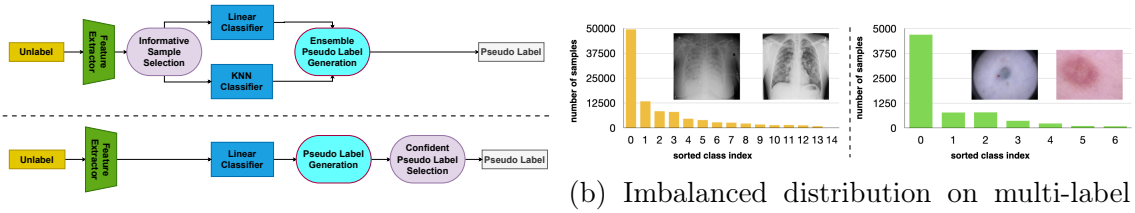
Chapter 4

ACPL: Anti-curriculum Pseudo-labelling for Semi-supervised Medical Image Classification

Abstract

Effective semi-supervised learning (SSL) in medical image analysis (MIA) must address two challenges: 1) work effectively on both multi-class (e.g., lesion classification) and multi-label (e.g., multiple-disease diagnosis) problems, and 2) handle imbalanced learning (because of the high variance in disease prevalence). One strategy to explore in SSL MIA is based on the pseudo labelling strategy, but it has a few shortcomings. Pseudo-labelling has in general lower accuracy than consistency learning, it is not specifically design for both multi-class and multi-label problems, and it can be challenged by imbalanced learning. In this paper, unlike traditional methods that select confident pseudo label by threshold, we propose a new SSL algorithm, called *anti-curriculum pseudo-labelling (ACPL)*, which introduces novel techniques to select informative unlabelled samples, improving training balance and allowing the model to work for both multi-label and multi-class problems, and to estimate pseudo labels by an accurate ensemble of classifiers (improving pseudo label accuracy). We run extensive experiments to evaluate ACPL on two public medical image classification benchmarks: Chest X-Ray14 for thorax disease multi-label classification and ISIC2018 for skin lesion multi-class classification. Our method outperforms previous SOTA SSL methods on both datasets[1][2].

[1]Supported by Australian Research Council through grants DP180103232 and FT190100525.
[2]Code is available at https://github.com/FBLADL/ACPL



(a) Diagram of our ACPL (top) and traditional pseudo-label SSL (bottom)

(b) Imbalanced distribution on multi-label Chest X-ray14 [192] (left) and multi-class ISIC2018 [182] (right)

Figure 4.1: In (a), we show diagrams of the proposed ACPL (top) and the traditional pseudo-label SSL (bottom) methods, and (b) displays histograms of images per label for the multi-label Chest X-ray14 [192] (left) and multi-class ISIC2018 [182] (right).

4.1 Introduction

Deep learning has shown outstanding results in medical image analysis (MIA) [113, 178, 179]. Compared to computer vision, the labelling of MIA training sets by medical experts is significantly more expensive, resulting in low availability of labelled images, but the high availability of unlabelled images from clinics and hospitals databases can be explored in the modelling of deep learning classifiers. Furthermore, differently from computer vision problems that tend to be mostly multi-class and balanced, MIA has a number of multi-class (e.g., a lesion image of a single class) and multi-label (e.g., an image from a patient can contain multiple diseases) problems, where both problems usually contain severe class imbalances because of the variable prevalence of diseases (see Fig. 4.1-(b)). Hence, MIA semi-supervised learning (SSL) methods need to be flexible enough to work with multi-label and multi-class problems, in addition to handle imbalanced learning.

State-of-the-art (SOTA) SSL approaches are usually based on the consistency learning of unlabelled data [11, 12, 170] and self-supervised pre-training [117]. Even though consistency-based methods show SOTA results on multi-class SSL problems, pseudo-labelling methods have shown better results for multi-label SSL problems [156]. Pseudo-labelling methods provide labels to confidently classified unlabelled samples that are used to re-train the model [103]. One issue with pseudo-labelling SSL methods is that the confidently classified unlabelled samples represent the least informative ones [161] that, for imbalanced problems, are likely to belong to the majority classes. Hence, this will bias the classification toward the majority classes and most likely deteriorate the classification accuracy of the minority classes. Also, selecting confident pseudo-labelled samples is challenging in multi-class, but even more so in multi-label problems. Previous papers [4, 156] use a fixed threshold for all classes, but a class-wise threshold that addresses imbalanced learning and correlations between classes in multi-label problems would enable more accurate pseudo-label predictions. However, such class-wise thresh-

old is hard to estimate without knowing the class distributions or if we are dealing with a multi-class or multi-label problem. Furthermore, using the model output for the pseudo-labelling process can also cause confirmation bias [2], whereby the assignment of incorrect pseudo-labels will increase the model confidence in those incorrect predictions, and consequently decrease the model accuracy.

In this paper, we propose the **anti-curriculum pseudo-labelling (ACPL)**, which addresses multi-class and multi-label imbalanced learning SSL MIA problems. First, we introduce a new approach to select the most informative unlabelled images to be pseudo-labelled. This is motivated by our argument that there exists a distribution shift between unlabelled and labelled samples for SSL. An effective learning curriculum must focus on informative unlabelled samples that are located as far as possible from the distribution of labelled samples. As a result, these informative samples are likely to belong to the minority classes in MIA imbalanced learning problems. Selecting these informative samples will naturally balance the training process and, given that they are selected before the pseudo-labelling process, we eliminate the need for estimating a class-wise classification threshold, facilitating our model to work well on multi-class and multi-label problems. The information content measure of an unlabelled sample is computed with our proposed cross-distribution sample informativeness that outputs how close an unlabelled sample is from the set of labelled anchor samples (anchor samples are highly informative labelled samples). Second, we introduce a new pseudo-labelling mechanism, called informative mixup, which combines the model classification with a K-nearest neighbor (KNN) classification guided by sample informativeness to improve prediction accuracy and mitigate confirmation bias. Third, we propose the anchor set purification method that selects the most informative pseudo-labelled samples to be included in the labelled anchor set to improve the pseudo-labelling accuracy of the KNN classifier in later training stages.

To summarise, our ACPL approach selects highly informative samples for pseudo-labelling (addressing MIA imbalanced classification problems and allowing multi-label multi-class modelling) and uses an ensemble of classifiers to produce accurate pseudo labels (tackling confirmation bias to improve classification accuracy), where the main technical contributions are:

- A novel information content measure to select informative unlabelled samples named **cross-distribution sample informativeness**;

- A new pseudo-labelling mechanism, called **informative mixup**, which generates pseudo labels from an ensemble of deep learning and KNN classifiers; and

- A novel method, called **anchor set purification (ASP)**, to select informative pseudo-labelled samples to be included in the labelled anchor set to improve the pseudo-labelling accuracy of the KNN classifier.

We evaluate ACPL on two publicly available medical image classification datasets, namely the Chest X-Ray14 for thorax disease multi-label classification [192] and the ISIC2018 for skin lesion multi-class classification [29, 182]. Our method outperforms the current SOTA methods in both datasets.

4.2 Related Work

We first review consistency-based and pseudo-labelling SSL methods. Then, we discuss the curriculum and anti-curriculum learning literature for fully and semi-supervised learning and present relevant SSL MIA methods.

Consistency-based SSL optimises the classification prediction of labelled images and minimises the prediction outputs of different views of unlabelled images, where these views are obtained from different types of image perturbations, such as spatial/temporal [99, 177], adversarial [136], or data augmentation [11, 12, 170]. The performance of the consistency-based methods can be further improved with self-supervised pre-training [117]. Even though consistency-based SSL methods show SOTA results in many benchmarks [170], they depend on a careful design of perturbation functions that requires domain knowledge and would need to be adapted to each new type of medical imaging. Furthermore, Rizve et al. [156] show that pseudo-labelling SSL methods are more accurate for multi-label problems.

Pseudo-labelling SSL methods [19, 156, 163, 214] train a model with the available labelled data, estimate the pseudo labels of unlabelled samples classified with high confidence [103], then take these pseudo-labelled samples to re-train the model. As mentioned above in Sec. 4.1 pseudo-label SSL approaches can bias classification toward the majority classes in imbalanced problems, is not seamlessly adaptable to multi-class and multi-label problems, and can also lead to confirmation bias. We argue that the improvement of pseudo-labelling SSL methods depends on the selection of informative unlabelled samples to address the majority class bias and the adaptation to multi-class and multi-label problems, and an accurate pseudo-labelling mechanism to handle confirmation bias, which are two points that we target with this paper. The selection of training samples based on their information content has been studied by fully supervised **curriculum and anti-curriculum learning methods** [206]. Curriculum learning focuses on the easy samples in the early training stages and gradually includes the hard samples in the later training stages, where easy samples [9, 80, 98] are usually defined as samples that have small losses during training, and hard samples tend to have large losses. On the other hand, anti-curriculum focuses of the hard samples first and transitions to the easy samples later in the training [78, 89]. The methods above have been designed to work in fully supervised learning.

Since we target accurate SSL of imbalanced multi-class and multi-label methods, we follow anti-curriculum learning that pseudo-labels the most informative samples

which are likely to belong to the minority classes (consequently, helping to balance the training) and enable the selection of samples without requiring the estimation of a class-wise classification threshold (enabling a seamless adaptation to multi-class and multi-label problems).

The main benchmarks for SSL in MIA study the multi-label classification of chest X-ray (CXR) images [76, 192] and multi-class classification of skin lesions [29, 182]. For **CXR SSL** classification, pseudo-labelling methods have been explored [4], but SOTA results are achieved with consistency learning approaches [35, 111, 117, 118, 184]. For **skin lesion SSL** classification, the current SOTA is also based on consistency learning [118], with pseudo-labelling approaches [7] not being competitive. We show that our proposed pseudo-labelling method ACPL can surpass the consistency-based SOTA on both benchmarks, demonstrating the value of selecting highly informative samples for pseudo labelling and of the accurate pseudo labels from the ensemble of classifiers. We also show that our ACPL improves the current computer vision SOTA [156] applied to MIA, demonstrating the limitation of computer vision methods in MIA and also the potential of our approach to be applied in more general SSL problems.

4.3 Methods

To introduce our SSL method ACPL, assume that we have a small labelled training set $\mathcal{D}_L = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{D}_L|}$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^{H \times W \times C}$ is the input image of size $H \times W$ with C colour channels, and $\mathbf{y}_i \in \{0,1\}^{|\mathcal{Y}|}$ is the label with the set of classes denoted by $\mathcal{Y} = \{1, ..., |\mathcal{Y}|\}$ (note that $\mathbf{y}_i$ is a one-hot vector for multi-class problems and a binary vector in multi-label problems). A large unlabelled training set $\mathcal{D}_U = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}_U|}$ is also provided, with $|\mathcal{D}_L| << |\mathcal{D}_U|$. We assume the samples from both datasets are drawn from the same (latent) distribution. Our algorithm also relies on the pseudo-labelled set $\mathcal{D}_S$ that is composed of pseudo-labelled samples classified as informative unlabelled samples, and an anchor set $\mathcal{D}_A$ that contains informative pseudo-labelled samples. The goal of ACPL is to learn a model $p_\theta : \mathcal{X} \rightarrow [0,1]^{|\mathcal{Y}|}$ parameterised by θ using the labelled, unlabelled, pseudo-labelled, and anchor datasets.

Below, in Sec. 4.3.1, we introduce our ACPL optimisation that produces accurate pseudo labels to unlabelled samples following an anti-curriculum strategy, where highly informative unlabelled samples are selected to be pseudo-labelled at each training stage. In Sec. 4.3.2, we present the information criterion of an unlabelled sample, referred to as *cross distribution sample informativeness (CDSI)*, based on the dissimilarity between the unlabelled sample and samples in the anchor set $\mathcal{D}_A$. The pseudo labels for the informative unlabelled samples are generated using the proposed *informative mixup (IM)* method (Sec. 4.3.3) that mixes up the results from the model $p_\theta(.)$ and a K nearest neighbor (KNN) classifier using the anchor set. At the end of each training stage, the

Algorithm 1 Anti-curriculum Pseudo-labelling Algorithm

require: Labelled set $\mathcal{D}_L$, unlabelled set $\mathcal{D}_U$, and number of training stages T
initialise $\mathcal{D}_A = \mathcal{D}_L$, and $t = 0$
warm-up train $p_{\theta_t}(\mathbf{x})$ **with**
$\theta_t = \arg\min_\theta \frac{1}{|\mathcal{D}_L|} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_L} \ell(\mathbf{y}_i, p_\theta(\mathbf{x}_i))$
while $t < T$ **or** $|\mathcal{D}_U| \neq 0$ **do**
  build pseudo-labelled dataset using CDSI from (4.2) **and IM from** (4.6)**:**

$$\mathcal{D}_S = \{(\mathbf{x}, \tilde{\mathbf{y}}) | \mathbf{x} \in \mathcal{D}_U, h(f_{\theta_t}(\mathbf{x}), \mathcal{D}_A) = 1, \\ \tilde{\mathbf{y}} = g(f_{\theta_t}(\mathbf{x}), \mathcal{D}_A)\}$$

  update anchor set with ASP from (4.7)**:**

$$\mathcal{D}_A = \mathcal{D}_A \bigcup (\mathbf{x}, \tilde{\mathbf{y}}), \text{ where} \\ (\mathbf{x}, \tilde{\mathbf{y}}) \in \mathcal{D}_S, \text{ and } a(f_{\theta_t}(\mathbf{x}), \mathcal{D}_U, \mathcal{D}_A) = 1$$

  $t \leftarrow t + 1$
  optimise (4.1) **using** $\mathcal{D}_L, \mathcal{D}_S$ **to obtain** $p_{\theta_t}(\mathbf{x})$
  update labelled and unlabelled sets:
  $\mathcal{D}_L \leftarrow \mathcal{D}_L \bigcup \mathcal{D}_S, \mathcal{D}_U \leftarrow \mathcal{D}_U \setminus \mathcal{D}_S$
end while
return $p_{\theta_t}(\mathbf{x})$

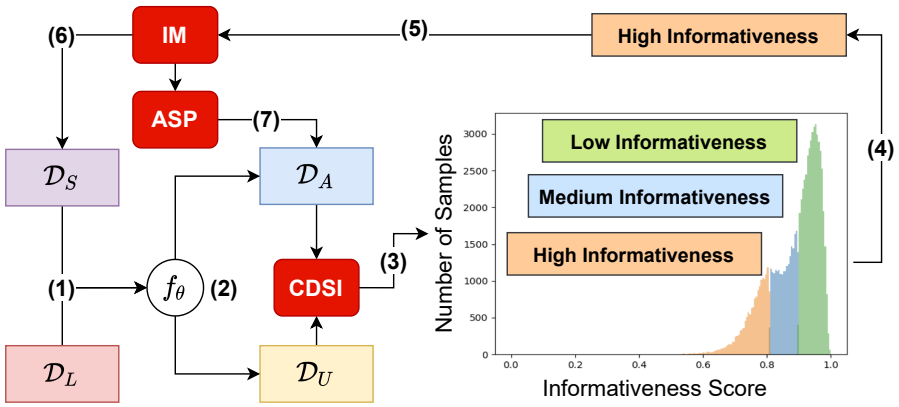


Figure 4.2: Anti-curriculum pseudo-labelling (ACPL) algorithm. The algorithm is divided into the following iterative steps: 1) train the model with $\mathcal{D}_S$ and $\mathcal{D}_L$; 2) extract the features from the anchor and unlabelled samples; 3) estimate information content of unlabelled samples with CDSI from (4.4) with anchor set $\mathcal{D}_A$; 4) partition the unlabelled samples into high, medium and low information content using (4.2); 5) assign a pseudo label to high information content unlabelled samples with IM from (4.6); 6) update $\mathcal{D}_S$ with new pseudo-labelled samples; and 7) update $\mathcal{D}_A$ with ASP in (4.7).

anchor set is updated with the *anchor set purification (ASP)* method (Sec. 4.3.4) that only keeps the most informative subset of pseudo-labelled samples, according to the *CDSI* criterion.

4.3.1 ACPL Optimisation

Our ACPL optimisation, described in Alg. 1 and depicted by Fig. 4.2, starts with a warm-up supervised training of the parameters of the model $p_\theta(.)$ using only the labelled set $\mathcal{D}_L$. For the rest of the training, we use the sets of labelled and unlabelled samples, $\mathcal{D}_L$ and $\mathcal{D}_U$, and update the pseudo-labelled set $\mathcal{D}_S$ and the anchor set $\mathcal{D}_A$ containing the informative unlabelled and pseudo-labelled samples, where $\mathcal{D}_S$ start as an empty set and $\mathcal{D}_A$ starts with the samples in $\mathcal{D}_L$. The optimisation iteratively minimises the following cost function:

$$\begin{aligned}\ell_{ACPL}(\theta, \mathcal{D}_L, \mathcal{D}_S) = &\frac{1}{|\mathcal{D}_L|} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_L} \ell(\mathbf{y}_i, p_\theta(\mathbf{x}_i)) \\ &+ \frac{1}{|\mathcal{D}_S|} \sum_{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \mathcal{D}_S} \ell(\tilde{\mathbf{y}}_i, p_\theta(\mathbf{x}_i)),\end{aligned} \tag{4.1}$$

where $\ell(.)$ denotes a classification loss (*e.g.*, cross-entropy), θ is the model parameter, $\mathbf{y}_i$ is the ground truth, and $\tilde{\mathbf{y}}_i$ is the estimated pseudo label. After optimising (4.1),

the labelled and unlabelled sets are updated as $\mathcal{D}_L = \mathcal{D}_L \bigcup \mathcal{D}_S$ and $\mathcal{D}_U = \mathcal{D}_U \setminus \mathcal{D}_S$, and a new iteration of optimisation takes place.

4.3.2 Cross Distribution Sample Informativeness (CDSI)

The function that estimates if an unlabelled sample has high information content is defined by

$$h(f_\theta(\mathbf{x}), \mathcal{D}_A) = \begin{cases} 1, & p_\gamma(\zeta = \text{high}|\mathbf{x}, \mathcal{D}_A) > \tau, \\ 0, & \text{otherwise}, \end{cases} \tag{4.2}$$

where $\zeta \in \mathcal{Z} = \{\text{low}, \text{medium}, \text{high}\}$ represents the information content random variable, $\gamma = \{\mu_\zeta, \Sigma_\zeta, \pi_\zeta\}_{\zeta \in \mathcal{Z}}$ denotes the parameters of the Gaussian Mixture Model (GMM) $p_\gamma(.)$, and $\tau = \max\{p_\gamma(\zeta = \text{low}|\mathbf{x}, \mathcal{D}_A), p_\gamma(\zeta = \text{medium}|\mathbf{x}, \mathcal{D}_A)\}$. The function $p_\gamma(\zeta|\mathbf{x}, \mathcal{D}_A)$ can be decomposed into $p_\gamma(\mathbf{x}|\zeta, \mathcal{D}_A)p_\gamma(\zeta|\mathcal{D}_A)/p_\gamma(\mathbf{x}|\mathcal{D}_A)$, where

$$p_\gamma(\mathbf{x}|\zeta, \mathcal{D}_A) = n(d(f_\theta(\mathbf{x}), \mathcal{D}_A)|\mu_\zeta, \Sigma_\zeta), \tag{4.3}$$

with $n(.; \mu_\zeta, \Sigma_\zeta)$ denoting a Gaussian function with mean μ_ζ and covariance Σ_ζ, $p_\gamma(\zeta|\mathcal{D}_A) = \pi_\zeta$ representing the ownership probability of ζ (i.e., the weight of mixture ζ), and $p_\gamma(\mathbf{x}|\mathcal{D}_A)$ being a normalisation factor. The probability in (4.3) is computed with the density of the unlabelled sample $\mathbf{x}$ with respect to the anchor set $\mathcal{D}_A$, as follows:

$$d(f_\theta(\mathbf{x}), \mathcal{D}_A) = \frac{1}{K} \sum_{\substack{(f_\theta(\mathbf{x}_A), \mathbf{y}_A) \in \\ \mathcal{N}(f_\theta(\mathbf{x}), \mathcal{D}_A)}} \frac{f_\theta(\mathbf{x})^\top f_\theta(\mathbf{x}_A)}{\|f_\theta(\mathbf{x})\|_2 \|f_\theta(\mathbf{x}_A)\|_2}, \tag{4.4}$$

where $\mathcal{N}(f_\theta(\mathbf{x}), \mathcal{D}_A)$ represents the set of K-nearest neighbors (KNN) from the anchor set $\mathcal{D}_A$ to the input image feature $f_\theta(\mathbf{x})$, with each element in the set $\mathcal{D}_A$ denoted by $(f_\theta(\mathbf{x}_A), \mathbf{y}_A)$. The $F-$dimensional input image feature is extracted with $f_\theta : \mathcal{X} \to \mathbb{R}^F$ from the model $p_\theta(.)$ with $p_\theta(\mathbf{x}) = \sigma(f_\theta(\mathbf{x}))$, where $\sigma(.)$ is the final activation function to produce an output in $[0, 1]^{|\mathcal{Y}|}$. The parameters γ in (4.2) are estimated with the expectation-maximisation (EM) algorithm [39], every time after the anchor set is updated.

4.3.3 Informative Mixup (IM)

After selecting informative unlabelled samples with (4.2), we aim to produce reliable pseudo labels for them. We can provide two pseudo labels for each unlabelled sample $\mathbf{x} \in \mathcal{D}_U$: the model prediction from $p_\theta(\mathbf{x})$, and the K-nearest neighbor (KNN) prediction using the anchor set, as follows:

$$\begin{aligned} \tilde{\mathbf{y}}_{\text{model}}(\mathbf{x}) &= p_\theta(\mathbf{x}), \\ \tilde{\mathbf{y}}_{\text{KNN}}(\mathbf{x}) &= \frac{1}{K} \sum_{(f_\theta(\mathbf{x}_A), \mathbf{y}_A) \in \mathcal{N}(f_\theta(\mathbf{x}), \mathcal{D}_A)} \mathbf{y}_A. \end{aligned} \tag{4.5}$$

$\mathbf{y}_A$ is the label of anchor set samples. However, using any of the pseudo labels from (4.5) can be problematic for model training. The pseudo label in $\tilde{\mathbf{y}}_{\text{model}}(\mathbf{x})$ can cause confirmation bias, and the reliability of $\tilde{\mathbf{y}}_{\text{KNN}}(\mathbf{x})$ depends on the size and representativeness of the initial labelled set to produce accurate classification. Inspired by MixUp [229], we propose the **informative mixup** method that constructs the pseudo-labelling function $g(.)$ in (4.1) with a linear combination of $\tilde{\mathbf{y}}_{\text{model}}(\mathbf{x})$ and $\tilde{\mathbf{y}}_{\text{KNN}}(\mathbf{x})$ weighted by the density score from (4.4), as follows:

$$\begin{aligned}\tilde{\mathbf{y}} = g(f_\theta(\mathbf{x}), \mathcal{D}_A) = d(f_\theta(\mathbf{x}), \mathcal{D}_A) \times \tilde{\mathbf{y}}_{\text{model}}(\mathbf{x}) \\ + (1 - d(f_\theta(\mathbf{x}), \mathcal{D}_A)) \times \tilde{\mathbf{y}}_{\text{KNN}}(\mathbf{x}).\end{aligned} \tag{4.6}$$

The informative mixup in (4.6) is different from MixUp [229] because it combines the classification results of the same image from two models instead of the classification from the same model of two images. Furthermore, our informative mixup weights the the classifiers with the density score to reflect the trade-off between $\tilde{\mathbf{y}}_{\text{model}}(\mathbf{x})$ and $\tilde{\mathbf{y}}_{\text{KNN}}(\mathbf{x})$. Since informative samples are selected from a region of the anchor set with low feature density, the KNN prediction $\tilde{\mathbf{y}}_{\text{KNN}}(\mathbf{x})$ is less reliable than $\tilde{\mathbf{y}}_{\text{model}}(\mathbf{x})$, so by default, we should trust more the model classification. The weighting between the two predictions in (4.6) reflects this observation, where $\tilde{\mathbf{y}}_{\text{model}}(\mathbf{x})$ will tend have a larger weight given that $d(f_\theta(\mathbf{x}), \mathcal{D}_A)$ is usually larger than 0.5, as displayed in Fig. 4.2 (see the informativeness score histogram at the bottom-right corner). When the sample is located in a high-density region, we place most of the weight on the model prediction given that in such case, the model is highly reliable. On the other hand, when the sample is in a low-density region, we try to balance a bit more the contribution of both the model and KNN predictions, given the low reliability of the model.

4.3.4 Anchor Set Purification (ASP)

After estimating the pseudo label for informative unlabelled samples, we aim to update the anchor set with informative pseudo-labelled samples to maintain density score from (4.4) accurate in later training stages. However, adding all pseudo-labelled samples will cause anchor set over-sized and increase hyper-parameter sensitivity. Thus, we propose the Anchor Set Purification (ASP) module to select the least connected pseudo-labelled samples to be inserted in the anchor set, as in (see Fig. 4.3):

$$a(f_\theta(\mathbf{x}), \mathcal{D}_U, \mathcal{D}_A) = \begin{cases} 1, & c(f_\theta(\mathbf{x}), \mathcal{D}_U, \mathcal{D}_A) \leq \alpha, \\ 0, & \text{otherwise}, \end{cases} \tag{4.7}$$

where the pseudo-labelled samples with $a(f_\theta(\mathbf{x}), \mathcal{D}_U, \mathcal{D}_A) = 1$ and $\tilde{\mathbf{y}} = g(f_\theta(\mathbf{x}), \mathcal{D}_A)$ from (4.6) are inserted into the anchor set. The information content $c(f_\theta(\mathbf{x}), \mathcal{D}_U, \mathcal{D}_A)$ of a pseudo-labelled sample $f_\theta(\mathbf{x})$ in (4.7) is computed in three steps (see Fig. 4.3): 1) find the KNN samples $\mathcal{N}(f_\theta(\mathbf{x}), \mathcal{D}_A)$ from $f_\theta(\mathbf{x})$ to the anchor set $\mathcal{D}_A$; 2) for each

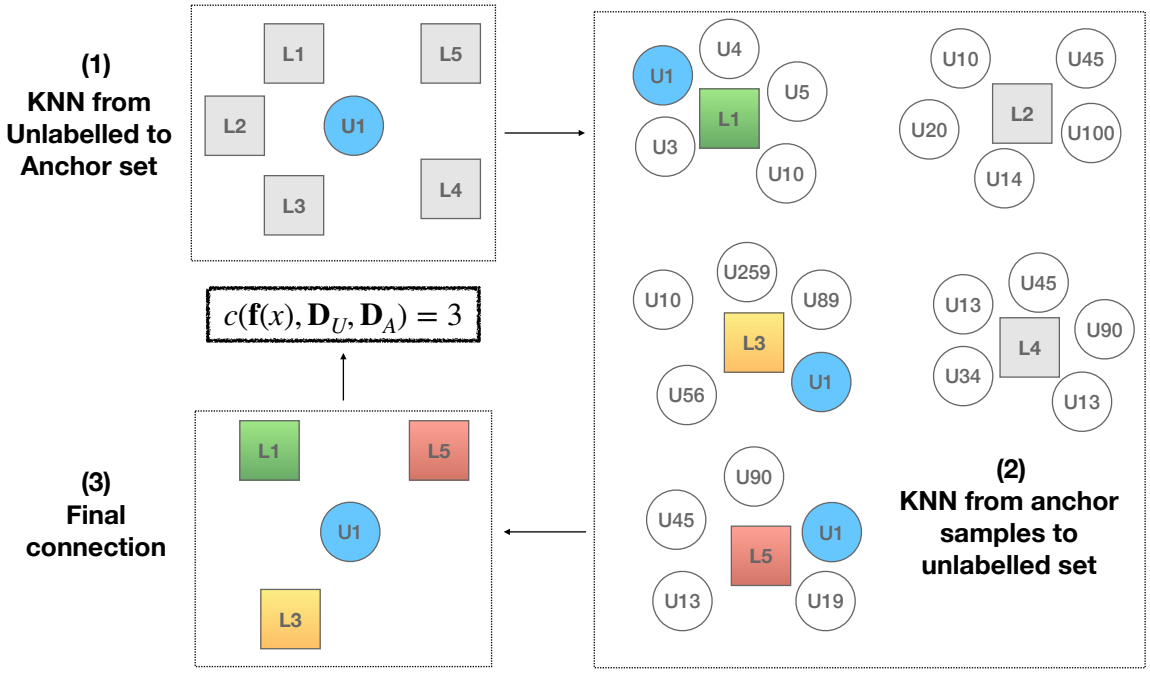


Figure 4.3: **ASP**: 1) find KNN samples from an informative unlabelled sample to the anchor set $\mathcal{D}_A$; 2) find KNN samples from each anchor sample of (1) to the unlabelled set $\mathcal{D}_U$; and 3) calculate the number of surviving nearest neighbours. Samples with the smallest values of $c(.)$ are selected to be inserted into $\mathcal{D}_A$.

Table 4.1: Mean AUC testing set results over the 14 disease classes of Chest X-Ray14 for different labelled set training percentages. * indicates the methods that use DenseNet-169 as backbone architecture. **Bold** number means the best result per label percentage and underline shows previous best results.

Method Type	Label Percentage	2%	5%	10%	15%	20%
Consistency based	SRC-MT* [118]	66.95	72.29	75.28	77.76	79.23
	NoTeacher [184]	72.60	77.04	77.61	N/A	79.49
	S^2MTS^2 [117]	74.69	78.96	79.90	80.31	81.06
Pseudo Label	Graph XNet* [4]	53.00	58.00	63.00	68.00	78.00
	UPS [156]	65.51	73.18	76.84	78.90	79.92
	Ours	**74.82**	**79.20**	**80.40**	**81.06**	**81.77**

of the K elements $(\mathbf{x}_A, \mathbf{y}_A) \in \mathcal{N}(f_\theta(\mathbf{x}), \mathcal{D}_A)$, find the KNN set $\mathcal{N}(f_\theta(\mathbf{x}_A), \mathcal{D}_U)$ from $f_\theta(\mathbf{x}_A)$ to the unlabelled set $\mathcal{D}_U$; and 3) $c(f_\theta(\mathbf{x}), \mathcal{D}_U, \mathcal{D}_A)$ is calculated to be the number of times that the pseudo-labelled sample $\mathbf{x}$ appears in the KNN sets $\mathcal{N}(f_\theta(\mathbf{x}_A), \mathcal{D}_U)$ for the K elements of set $\mathcal{N}(f_\theta(\mathbf{x}), \mathcal{D}_A)$. The threshold α in (4.7) is computed with $\alpha = \min_{\mathbf{x} \in \mathcal{D}_S} c(f_\theta(\mathbf{x}), \mathcal{D}_U, \mathcal{D}_A)$.

4.4 Experiments

For the experiments below, we use the Chest X-Ray14 [192] and ISIC2018 [29, 182] datasets.

Chest X-Ray14 contains 112,120 CXR images from 30,805 different patients. There are 14 labels (each label is a disease) and No Finding class, where each patient can have multiple labels, forming a multi-label classification problem. To compare with previous papers [4, 118], we adopt the official train/test data split [192]. We report the classification result on the test set (26K samples) using area under the ROC curve (AUC), and the learning process uses training sets containing different proportions of the labelled data in $\{2\%, 5\%, 10\%, 15\%, 20\%\}$.

ISIC2018 is a skin lesion dataset that contains 10,015 images with seven labels. Each image is associated with one of the labels, forming a multi-class classification problem. We follow the train/test split from [118] for fair comparison, where the training set contains 20% of labelled samples and 80% of unlabelled samples. We report the AUC, Sensitivity, and F1 score results.

4.4.1 Implementation Details

For both datasets, we use DenseNet-121 [72] as our backbone model. For Chest X-Ray14, the dataset pre-processing consists of resizing the images to 512 $\times$ 512 for faster processing. For the optimisation, we use Adam optimizer [93], batch size 16 and

Table 4.2: Class-level AUC testing set results comparison between our approach and other semi-supervised SOTA approaches trained with **20%** of labelled data on Chest Xray-14. * denotes the models use DenseNet-169 as backbone. **Bold** number means the best result per class and underlined shows second best results.

Method Type	Supervised	Consistency based			Pseudo-labelling		
Method	Densenet-121	MT [177] *	SRC-MT [118] *	S^2MTS^2 [117]	GraphXNet [4]	UPS [156]	Ours
Atelectasis	75.75	75.12	75.38	78.57	71.89	77.09	**79.53**
Cardiomegaly	80.71	87.37	87.7	88.08	87.99	85.73	**89.03**
Effusion	79.87	80.81	81.58	82.87	79.2	81.35	**83.56**
Infiltration	69.16	70.67	70.4	70.68	**72.05**	70.82	71.40
Mass	78.40	77.72	78.03	**82.57**	80.9	81.82	82.49
Nodule	74.49	73.27	73.64	76.60	71.13	76.34	**77.73**
Pneumonia	69.55	69.17	69.27	72.25	**76.64**	70.96	73.86
Pneumothorax	84.70	85.63	86.12	86.55	83.7	85.86	**86.95**
Consolidation	71.85	72.51	73.11	75.47	73.36	74.35	**75.50**
Edema	81.61	82.72	82.94	84.83	80.2	83.56	**84.95**
Emphysema	89.75	88.16	88.98	91.88	84.07	91.00	**93.36**
Fibrosis	79.30	78.24	79.22	81.73	80.34	80.87	**81.86**
Pleural Thicken	73.46	74.43	75.63	76.86	75.7	75.55	**77.60**
Hernia	86.05	**87.74**	87.27	85.98	87.22	85.62	85.89
Mean	78.19	78.83	79.23	81.06	78.00	79.92	**81.77**

learning rate 0.05. During training, we use data augmentation based on random crop and resize, and random horizontal flip. We first train 20 epochs on the initial labelled subset to warm-up the model for feature extraction. Then we train for 50 epochs, where in every 10 epochs we update the anchor set with ASP from Sec. 4.3.4. For the KNN classifier in (4.2), we set K to be 200 for 2% and 5% (of labelled data) and 50 for remaining label proportions. These values are set based on validation results, but our approach is robust to a large range K values – we show an ablation study that compares the performance of our method for different values of K. For ISIC2018, we resize the image to 224×224 for fair comparison with baselines. For the optimisation, we use Adam optimizer [93], batch size 32 and learning rate 0.001. During training, data augmentation is also based on random crop and resize, and random horizontal flip. We warm-up the model for 40 epochs and then we train for 100 epochs, where in every 20 epochs, we update the anchor set with ASP. For the KNN classifier, K is set to 100 based on validation set. The code is written in Pytorch [146] and we use two RTX 2080ti Gpus for all experiments. KNN computation takes 5 sec for Chest X-ray14 unlabelled samples with Faiss [85] library for faster processing. We follow [117, 118, 177] to maintain an exponential moving average (EMA) version of the trained model, which is only used for evaluation not for training.

4.4.2 Thorax Disease Classification Result

For the results on Chest X-Ray14 in Table 4.1, our method, NoTeacher [184], UPS [156], and S^2MTS^2 [117] use the DenseNet-121 backbone, while SRC-MT [118] and GraphXNet [4] use DenseNet-169 [72]. SRC-MT [118] is a consistency-based SSL; NoTeacher [184] extends MT by replacing the EMA process with two networks combined with a probabilistic graph model; S^2MTS^2 [117] combines self-supervised pre-training with MT fine-tuning; and GraphXNet [4] constructs a graph from dataset samples and assigns pseudo labels to unlabelled samples through label propagation; and UPS [156] applies probability and uncertainty thresholds to enable the pseudo labelling of unlabelled samples. All methods use the official test set [192]. Our approach achieves the SOTA results for all percentages of training labels. Compared to the pseudo-labelling approaches UPS and GraphXNet, our approach outperforms them by a margin between 3% to 20%. Compared to the consistency-based approaches SRC-MT and NoTeacher, our method consistently achieves 2% improvement for all cases, even though we use a backbone architecture of lower capacity (i.e., DenseNet-121 instead of DenseNet-169). Compared with the previous SOTA, our method outperforms S^2MTS^2 [117] by 1% AUC in all cases, which is remarkable because our method is initialised with an ImageNet pre-trained model instead of an expensive self-supervised pre-training approach.

The class-level performances using 20% of the labelled data of SSL methods are shown in Table 4.2, which demonstrates that our method achieves the best result in 10 out of the 14 classes. Our method surpasses the previous pseudo-labelling method GraphXNet by 3.7% and threshold based pseudo-labelling method [156] by 1.8%. Our method also outperforms consistency-based methods MT [177] and SRC-MT [118] by more than 2%. For method S^2MTS^2 [117] with self-supervised learning, our method can outperform it using an ImageNet pre-trained model, alleviating the need of a computationally expensive self-supervised pre-training.

4.4.3 Skin Lesion Classification Result

We show the results on ISIC2018 in Table 4.3, where competing methods are based on self-training [7], generative adversarial network (GAN) to augment the labelled set [42], temporal ensembling [99], MT [177] and its extension [118], and a DenseNet-121 [72] baseline trained with 20% of the training set. Compared with consistency-based approaches [111, 118, 177], our method improves between 0.7% and 3% in AUC and around 1% in F1 score. Our method also outperforms previous self-training approach [7] by a large margin in all measures.

Table 4.3: AUC, Sensitivity and F1 testing results on ISIC2018, where 20% of the training set is labelled. **Bold** shows the best result per measure, and underline shows second best results.

Method	AUC	Sensitivity	F1
Supervised	90.15	65.50	52.03
SS-DCGAN [42]	91.28	67.72	54.10
TCSE [111]	92.24	68.17	58.44
TE [99]	92.70	69.81	59.33
MT [177]	92.96	69.75	59.10
SRC-MT [118]	93.58	71.47	60.68
Self-training [7]	90.58	67.63	54.51
Ours	**94.36**	**72.14**	**62.23**

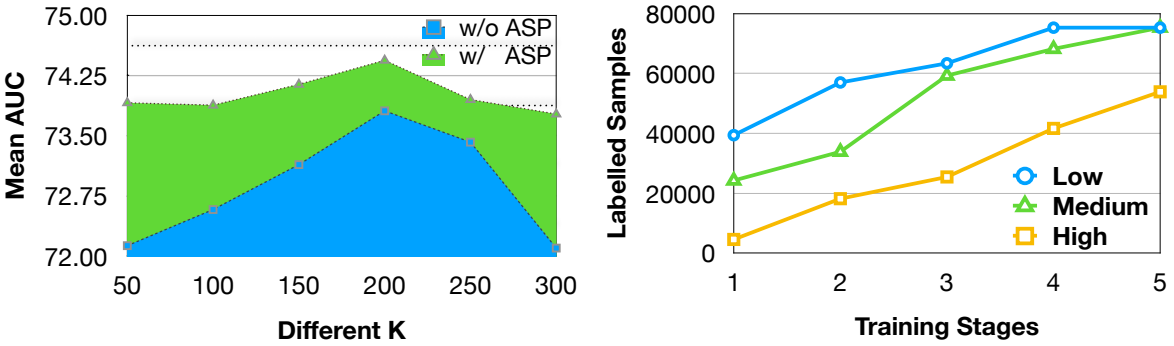


Figure 4.4: (Left) Mean AUC testing results for different values for K in the KNN (for CDSI in (4.4) and pseudo-labelling in (4.5)), where the green region uses ASP and blue region does not use ASP. (Right) Mean size of $\mathcal{D}_L$ at every training stage when adding unlabelled samples of high, medium and low information content according to (4.2). Model is trained on Chest X-Ray14, where 2% of the training is labelled.

4.4.4 Ablation Study

For the ablation study, we test each of our three contributions and visualize the data distribution of selected subset with high and low informative samples on the Chest X-Ray14 [192] with 2% labelled training set, where for CDSI and ASP, due to time and computation resources limited, we run each experiment three times and show the mean and standard deviation of the AUC results.

Cross-distribution Sample Informativeness (CDSI). We first study in Table 4.4 how performance is affected by pseudo-labelling unlabelled samples with different degrees of informativeness (low, medium and high) using our CDSI. Starting from the baseline classifier DenseNet-121 that reaches an AUC of 65%, we observe that pseudo-labelling low-information content unlabelled samples yields the worst result (around

Table 4.4: Ablation study on Chest X-ray14 (2% labelled). Starting with a baseline classifier (DenseNet-121), we test the selection of unlabelled samples (to be provided with a pseudo-label) with different information content, according to (4.2) (i.e., low, medium, high), and the use of the anchor set purification (ASP) module.

Information Content	ASP	AUC ± std
Baseline		65.84 ± 0.14
Low	✗	67.18 ± 2.40
	✓	67.76 ± 1.05
Medium	✗	70.83 ± 1.49
	✓	71.16 ± 0.51
High	✗	73.81 ± 0.75
	✓	**74.44 ± 0.38**

67% AUC) and selecting high-information content unlabelled samples produces the best result (around 73% AUC). Figure 4.4 (right) plots how the size of the labelled set $\mathcal{D}_L$ during training depends on the degree of informativeness of the unlabelled samples to be pseudo-labelled. These results show that: 1) unlabelled samples of high-information content enables the construction of a smaller labelled set (compared with unlabelled samples of low- or medium-information content), allowing a more efficient training process that produces a more accurate KNN classifier; and 2) the standard deviation of the results in Table 4.4 are smaller when selecting the unlabelled samples of high-information content, compared with the low- or medium-information content. This second point can be explained by the class imbalance issue in Chest X-Ray14, where the selection of low-information content samples will enable the training of majority classes, possibly producing an ineffective training for the minority classes that can increase the variance in the results.

Anchor Set Purification (ASP). Also in Table 4.4, we compare ASP with an alternative method that selects all pseudo-labelled samples to be included into the anchor set for the low-, medium- and high-information content unlabelled samples. Results show that the ASP module improves AUC between 0.3% and 1.0% and reduces standard deviation between 0.4% and 1.4%. This demonstrates empirically that the ASP module enables the formation of a more informative anchor set that improves the pseudo-labelling accuracy, and consequently the final AUC results. Furthermore, in Figure 4.4 (left), ASP is shown to stabilise the performance of the method with respect to $K \in \{50, 100, 150, 200, 250, 300\}$ for the KNN classifier of (4.4). In particular, with ASP, the difference between the best and worst AUC results is around 1%, while without ASP, the difference grows to 2%. This can be explained by the fact that without ASP, the anchor set grows quickly with relatively less informative pseudo-labelled samples, which reduces the stability of the method.

Table 4.5: AUC testing set results on Chest X-ray14 (2% labelled) for different pseudo labelling strategies (α denotes the linear coefficient combining the model and KNN predictions).

Pseudo-label Strategies	Methods	AUC
Baseline	-	65.84
Single Prediction	Model prediction	72.63
	KNN prediction	72.45
Mixup	random sampled α	73.23
	MixUp [229]	69.28
Ours	Informative Mixup	**74.44**

Informative Mixup (IM) In Table 4.5, we show that our proposed IM in (4.6) produces a more accurate pseudo-label, where we compare it with alternative pseudo-label methods, such as with only the model prediction, only the KNN prediction, random sample α from beta distribution to replace the density score in (4.6), and regular MixUp [229]. It is clear that the use of model or KNN predictions alone as pseudo labels has performance gap. This most likely because of confirmation bias (former case) or the inaccuracy of the KNN classifier (latter case). MixUp [229] does not show good accuracy either, as also observed in [189] and [90], when MixUp is performed in multi-label images or multiple single-object images. The random sampling of α for replacing density score shows a better result than MixUp, but the lack of an image-based weight to balance the two predictions, like in (4.6), damages performance. Our proposed IM shows a result that is at least 1% better than any of the other pseudo-labelling approaches, showing the importance of using the density of the unlabelled sample in the anchor set to weight the contribution of the model and KNN classifiers.

The **imbalanced learning mitigation** is studied in Figure 4.5, which shows the histogram of label distribution in percentage (for a subset of four disease minority classes and the No Finding majority class) by selecting unlabelled samples of high (blue) and low (yellow) information content. We also show the original label distribution in green for reference.

Notice that the selection of highly informative samples significantly increases the percentage of disease minority classes (from between 5% and 10% to almost 30%) and decreases the percentage of the No Finding majority class (from 60% to 30%), creating a more balanced distribution of these five classes. This indicates that our informative sample selection can help to mitigate the issue of imbalanced learning. We include the full 14-classes histograms in the Appendix A.

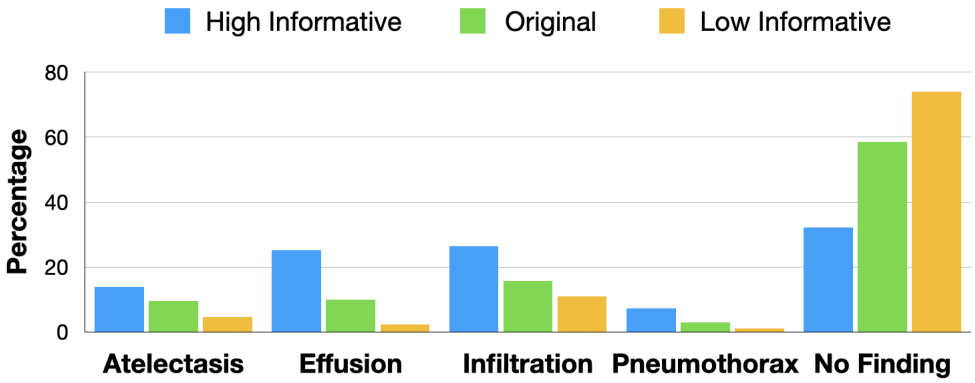


Figure 4.5: The selection of highly informative unlabelled samples (blue) promote a more balanced learning process, where the difference in the number of samples belonging to the minority or majority classes is smaller than if we selected unlabelled samples with low informativeness (yellow). Green shows the original data distribution]. Full 14-class distributions are shown in the Appendix A.

4.5 Discussion and Conclusion

In this work, we introduced the anti-curriculum pseudo-labelling (ACPL) SSL method. Unlike traditional pseudo-labelling methods that use a threshold to select confidently classified samples, ACPL uses a new mechanism to select highly informative unlabelled samples for pseudo-labelling and an ensemble of classifiers to produce accurate pseudo-labels. This enables ACPL to address MIA multi-class and multi-label imbalanced classification problems. We show in the experiments that ACPL outperforms previous consistency-based, pseudo-label based and self-supervised SSL methods in multi-label Chest X-ray14 and multi-class ISIC2018 benchmarks. We demonstrate in the ablation study the influence of each of our contributions and we also show how our new selection of informative samples addresses MIA imbalanced classification problems. For future work, it is conceivable that ACPL can be applied to more general computer vision problems, so we plan to test ACPL in traditional computer vision benchmarks. We would also explore semi-supervised classification with out-of-distribution (OOD) data in the initial labelled and unlabelled sets as our method currently assume all samples are in-distribution.

Statement of Authorship

Title of Paper	NVUM: Non-volatile Unbiased Memory for Robust Medical Image Classification
Publication Status	☐ Published ☑ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Published at Internation Conference on Medical Imaging Computing and Computer-Assisted Intervention (MICCAI), 2022, Early accept

Principal Author

Name of Principal Author (Candidate)	Fengbei Liu
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper and revision
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: 09/14/2023

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 in 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	Conducted experiments and wrote the revision
Signature	Date: 09/15/2023

Name of Co-Author	Yu Tian
Contribution to the Paper	Discussion and commented on the revision
Signature	Date

Name of Co-Author	Yuyuan Liu
Contribution to the Paper	Discussion and commented on the revision
Signature	Date

Name of Co-Author	Chong Wang
Contribution to the Paper	Discussion and commented on the revision
Signature	Date

Name of Co-Author	Vasileios Belagiannis
Contribution to the Paper	Discussion and commented on the revision
Signature	Date 19.09.2023

Name of Co-Author	Gustavo Carneiro
Contribution to the Paper	Discussion and wrote the revision
Signature	Date 22/09/2023

Chapter 5

NVUM: Non-volatile Unbiased Memory for Robust Medical Image Classification

Abstract

Real-world large-scale medical image analysis (MIA) datasets have three challenges: 1) they contain noisy-labelled samples that affect training convergence and generalisation, 2) they usually have an imbalanced distribution of samples per class, and 3) they normally comprise a multi-label problem, where samples can have multiple diagnoses. Current approaches are commonly trained to solve a subset of those problems, but we are unaware of methods that address the three problems simultaneously. In this paper, we propose a new training module called Non-Volatile Unbiased Memory (NVUM), which non-volatility stores running average of model logits for a new regularization loss on noisy multi-label problem. We further unbias the classification prediction in NVUM update for imbalanced learning problem. We run extensive experiments to evaluate NVUM on new benchmarks proposed by this paper, where training is performed on noisy multi-label imbalanced chest X-ray (CXR) training sets, formed by Chest-Xray14 and CheXpert, and the testing is performed on the clean multi-label CXR datasets OpenI and PadChest. Our method outperforms previous state-of-the-art CXR classifiers and previous methods that can deal with noisy labels on all evaluations. Our code is available at https://github.com/FBLADL/NVUM. [1]

[1]This work was supported by the Australian Research Council through grants DP180103232 and FT190100525.

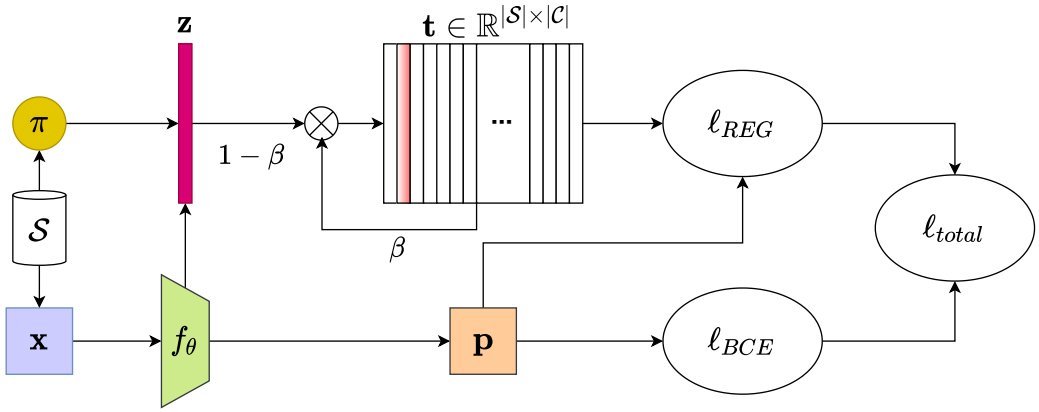


Figure 5.1: NVUM training algorithm: 1) sample input image $\mathbf{x}$ from training set $\mathcal{S}$ and calculate label distribution prior π; 2) train model f_θ and get sample logits $\mathbf{z}$ and prediction $\mathbf{p}$; 3) update memory $\mathbf{t}$ with (5.3); and 4) minimise the loss that comprises $\ell_{BCE}(.)$ in (5.1) and $\ell_{REG}(.)$ in (5.2).

5.1 Introduction and Background

The outstanding results shown by deep learning models in medical image analysis (MIA) [113, 116] depend on the availability of large-scale manually-labelled training sets, which is expensive to obtain. As a affordable alternative, these manually-labelled training sets can be replaced by datasets that are automatically labelled by natural language processing (NLP) tools that extract labels from the radiologists' reports [76, 192]. However, the use of these alternative labelling processes often produces unreliably labelled datasets because NLP-extracted disease labels, without verification by doctors, may contain incorrect labels, which are called *noisy labels* [143, 144]. Furthermore, differently from computer vision problems that tend to be multi-class with a balanced distribution of samples per class, MIA problems are usually multi-label (e.g, a disease sample can contain multiple diagnosis), with severe class imbalances because of the variable prevalence of diseases. Hence, robust MIA methods need to be flexible enough to work with *noisy multi-label* and *imbalanced* problems. State-of-the-art (SOTA) noisy-label learning approaches are usually based on noise-cleaning methods [60, 106, 119]. Han et al. [60] propose to use two DNNs and use their disagreements to reject noisy samples from the training process. Li et al. [106] rely on semi-supervised learning that treats samples classified as noisy as unlabelled samples. Other approaches estimate the label transition matrix [54, 210] to correct model prediction. Even though these methods show state-of-the-art (SOTA) results in noisy-label problems, they have issues with imbalanced and multi-label problems. First, noise-cleaning methods usually rely on detecting noisy samples by selecting large training loss samples, which are ei-

ther removed or re-labelled. However, in imbalanced learning problems, such training loss for clean-label training samples, belonging to minority classes, can be larger than the loss for noisy-label training samples belonging to majority classes, so these SOTA noisy-label learning approaches may inadvertently remove or re-label samples belonging to minority classes. Furthermore, in multi-label problems, the same sample can have a mix of clean and noisy labels, so it is hard to adapt SOTA noisy-label learning approaches to remove or re-label particular labels of each sample. Another issue in multi-label problems faced by transition matrix methods is that they are designed to work for multi-class problems, so their adaptation to multi-label problems will need to account for the correlation between the multiple labels. Hence, current noisy-label learning approaches have not been designed to solve all issues present in noisy multi-label imbalanced real-world datasets. Current imbalanced learning approaches are usually based on decoupling classifier and representation learning [88, 176]. For instance, Kang et al. [88] notice that learning with an imbalanced training set does not affect the representation learning, so they only adjust for imbalanced learning when training the classifier. Tang et al. [176] identify causal effect in stochastic gradient descent (SGD) momentum update on imbalanced datasets and propose a de-confounded training scheme. Another type of imbalanced learning is based on loss weighting [18, 175] that up-weights the minority classes [18] or down-weights the majority classes [175]. Furthermore, Menon et al. [135] discover that decoupling approach that based on correlation between classifier weight norm and data distribution is only applicable for SGD optimizer, which is problematic for MIA methods that tend to rely on other optimizers, such as Adam, that show better training convergence. Even though the papers above are effective for imbalanced learning problems, they do not consider the combination of imbalanced and noisy multi-label learning. To address the noisy multi-label imbalanced learning problems present in real-world MIA datasets, we introduce the **Non-volatile Unbiased Memory (NVUM)** training module, which is described in Fig. 5.1. Our contributions are:

- NVUM that stores a non-volatile running average of model logits to explore the multi-label noise robustness of the early learning stages. This memory module is used by a new regularisation loss to penalise differences between current and early-learning model logits;

- The NVUM update takes into account the class prior distribution to unbias the classification predictions estimated from the imbalanced training set;

- Two new noisy multi-label imbalanced evaluation benchmarks, where training is performed on chest X-ray (CXR) training sets from Chest Xray14 [192] and CheXpert [76], and testing is done on the clean multi-label CXR datasets OpenI [37] and PadChest [15].

5.2 Method

We assume the availability of a noisy-labelled training set $\mathcal{S} = \{(\mathbf{x}^i, \tilde{\mathbf{y}}^i)\}_{i=1}^{|\mathcal{S}|}$, where $\mathbf{x}^i \in \mathcal{X} \subset \mathbb{R}^{H\times W\times R}$ is the input image of size $H \times W$ with R colour channels, and $\tilde{\mathbf{y}}^i \in \{0,1\}^{|\mathcal{C}|}$ is the noisy label with the set of classes denoted by $\mathcal{C} = \{1, ..., |\mathcal{C}|\}$ (note that $\tilde{\mathbf{y}}_i$ represents a binary vector in multi-label problems, with each label representing one disease).

5.2.1 Non-volatile Unbiased Memory (NVUM) Training

To describe the NVUM training, we first need to define the model, parameterised by θ and represetned as a deep neural network, with $\mathbf{p} = \sigma(f_\theta(\mathbf{x}))$, where $\mathbf{p} \in [0,1]^{|\mathcal{C}|}$, $\sigma(.)$ denotes the sigmoid activation function and $\mathbf{z} = f_\theta(\mathbf{x})$, with $\mathbf{z} \in \mathcal{Z} \in \mathbb{R}^{|\mathcal{C}|}$ representing a logit. The training of the model $f_\theta(\mathbf{x})$ is achieved by minimising the following loss function:

$$\ell_{total}(\mathcal{S}, \mathbf{t}, \theta) = \frac{1}{|\mathcal{S}|} \sum_{(\mathbf{x}^i, \tilde{\mathbf{y}}^i)\in\mathcal{S}} \ell_{BCE}(\tilde{\mathbf{y}}^i, \mathbf{p}^i) + \ell_{REG}(\mathbf{t}^i, \mathbf{p}^i), \tag{5.1}$$

where ℓ_{BCE} denotes the binary cross-entropy loss for handling multi-label classification and ℓ_{REG} is a regularization term defined by:

$$\ell_{REG}(\mathbf{t}^i, \mathbf{p}^i) = \log(1 - \sigma((\mathbf{t}^i)^\top \mathbf{p}^i)). \tag{5.2}$$

here $\mathbf{t} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{C}|}$ is our proposed memory module designed to store an unbiased multi-label running average of the predicted logits for all training samples and $\mathbf{t}$ uses the class prior distribution π for updating, denoted by $\pi(c) = \frac{1}{|\mathcal{S}|}\sum_{i=1}^{|\mathcal{S}|} \tilde{\mathbf{y}}(c)$ for $c \in \{1, ..., |\mathcal{C}|\}$. The memory module $\mathbf{t}$ is initialised with zeros, as in $\mathbf{t}_0 = \mathbf{0}^{|\mathcal{S}|\times|\mathcal{C}|}$, and is updated in every epoch $k > 0$ with:

$$\mathbf{t}_k^i = \beta \mathbf{t}_{k-1}^i + (1-\beta)(\mathbf{z}_k^i - \log \pi), \tag{5.3}$$

where $\beta \in [0,1]$ is a hyper-parameter controlling the volatility of the memory storage, with β set to larger value representing a non-volatile memory and $\beta \approx 0$ denoting a volatile memory that is used in [63] for contrastive learning. To explore the early learning phenomenon, we set $\beta = 0.9$ so the regularization can enforce the consistency between the current model logits and the logits produced at the beginning of the training, when the model is robust to noisy label. Furthermore, to make the training robust to imbalanced problems, we subtract the log prior of the class distributions, which has the effect of increasing the logits with larger values for the classes with smaller prior. This counterbalances the issue faced by imbalanced learning problems, where the logits for the majority classes can overwhelm those from the minority classes, to the point that logit inconsistencies found by the regularization from noisy labels of

the majority classes may become indistinguishable from the clean labels from minority classes.
The effect of Eq. (5.2) can be interpreted by inspecting the loss gradient, which is proved in the Appendix B. The gradient of (5.1) is:

$$\begin{aligned} \nabla_\theta \ell_{total}(\mathcal{S}, \theta) &= \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbf{J}_{\mathbf{x}^i}(\theta)(\mathbf{p}^i - \tilde{\mathbf{y}}^i + \mathbf{g}^i), \\ \text{where} \quad \mathbf{g}_c^i &= -\sigma((\mathbf{t}^i)^\top(\mathbf{p}^i))\mathbf{p}_c^i(1 - \mathbf{p}_c^i)\mathbf{t}_c \end{aligned} \tag{5.4}$$

where $\mathbf{J}_{\mathbf{x}^i}(\theta)$ is the Jacobian matrix w.r.t. θ for the i^{th} sample. Assume $\mathbf{y}_c$ is the hidden true label of the sample $\mathbf{x}^i$, then the entry $\mathbf{t}_c^i > 0$ if $\mathbf{y}_c = 1$, and $\mathbf{t}_c^i < 0$ if $\mathbf{y}_c = 0$ at the early stages of training. During training, we consider four conditions explained below, where we assume that $\sigma((\mathbf{t}^i)^\top(\mathbf{p}^i))\mathbf{p}_c^i(1 - \mathbf{p}_c^i) > 0$. When the training sample has clean label:

- if $\tilde{\mathbf{y}}_c = \mathbf{y}_c = 1$, the gradient of the BCE term, $\mathbf{p}_c^i - \tilde{\mathbf{y}}_c^i \approx 0$ given that the model is likely to fit clean samples. With $\mathbf{t}_c^i > 0$, the sign of $\mathbf{g}_c^i$ is negative, and the model keeps training for these positive labels even after the early-training stages.
- if $\tilde{\mathbf{y}}_c = \mathbf{y}_c = 0$, the gradient of the BCE term, $\mathbf{p}_c^i - \tilde{\mathbf{y}}_c^i \approx 0$ given that the model is likely to fit clean samples. Given that $\mathbf{t}_c^i < 0$, we have $\mathbf{g}_c^i > 0$, and the model keeps training for these negative labels even after the early-training stages.

Therefore, adding $\mathbf{g}_c^i$ in total loss ensures that clean samples gradient magnitudes remains relatively high, encouraging a continuing optimisation using the clean label samples. For a noisy-label sample, we have:

- if $\tilde{\mathbf{y}}_c = 0$ and $\mathbf{y}_c = 1$, the gradient of the BCE loss is $\mathbf{p}_c^i - \tilde{\mathbf{y}}_c^i \approx 1$ because the model will not fit noisy label during early training stages. With $\mathbf{t}_c^i > 0$, we have $\mathbf{g}_c^i < 0$, which reduces the gradient magnitude from the BCE loss.
- if $\tilde{\mathbf{y}}_c = 1$ and $\mathbf{y}_c = 0$, $\mathbf{p}_c^i - \tilde{\mathbf{y}}_c^i \approx -1$. Given that $\mathbf{t}_c^i < 0$, we have $\mathbf{g}_c^i > 0$, which also reduces the gradient magnitude from the BCE loss.

Therefore, for noisy-label samples, $\mathbf{g}_c^i$ will counter balance the gradient from the BCE loss and diminish the effect of noisy-labelled samples in the training.

5.3 Experiment

Datasets. For the experiments below, we use the NIH Chest X-ray14 [192] and CheXpert [76] as noisy multi-label imbalanced datasets for training and Indiana OpenI [37] and PadChest [15] datasets for clean multi-label testing sets.

For the noisy sets, **NIH Chest X-ray14 (NIH)** contains 112,120 CXR images from 30,805 patients. There are 14 labels (each label is a disease), where each patient can have multiple diseases, forming a multi-label classification problem. For a fair comparison with previous papers [68, 152], we adopt the official train/test data split [192]. **CheXpert (CXP)** contains 220k images with 14 different diseases, and similarly to NIH, each patient can have multiple diseases. For pre-processing, we remove all lateral view images and treat uncertain and empty labels as negative labels. Given that the clean test set from CXP is not available and the clean validation set is too small for a fair comparison, we further split the training images into 90% training set and 10% *noisy* validation set with no patient overlapping. For the clean sets, **Indiana OpenI (OPI)** contains 7,470 frontal/lateral images with manual annotations. In our experiments, we only use 3,643 frontal view of images for evaluation. **PadChest (PDC)** is a large-scale dataset containing 158,626 images with 37.5% of images manually labelled. In our experiment, we only use the manually labelled samples as the clean test set. To keep the number of classes consistent between different datasets, we trim the training and testing sets based on the shared classes between these datasets [2].

Implementation Details. We use the ImageNet [158] pre-trained DenseNet121 [72] as the backbone model for $f_\theta(.)$ on NIH and CXP. We use Adam [93] optimizer with batch size 16 for NIH and 64 for CXP. For NIH, we train for 30 epochs with a learning rate of 0.05 and decay with 0.1 at 70% and 90% of the total of training epochs. Images are resized from 1024×1024 to 512×512 pixels. For data augmentation, we employ random resized crop and random horizontal flipping. For CXP, we train for 40 epochs with a learning rate of $1e^{-4}$ and follow the learning rate decay policy as on NIH. Images are resized to 224×224. For data augmentation, we employ random resized crop, 10 degree random rotation and random horizontal flipping. For both datasets, we use $\beta = 0.9$ and normalized by ImageNet mean and standard deviation.

All classification results are reported using area under the ROC curve (AUC). To report performance on clean test sets OPI and PDC, we adopt a common noisy label setup [60, 106] that selects the best performance checkpoint on noisy validation, which is the noisy test set of NIH and the noisy validation set of CXP. All experiments are implemented with Pytorch [147] and conducted on an NVIDIA RTX 2080ti GPU. The training takes 15 hours on NIH and 14 hours on CXP.

5.3.1 Experiments and Results

Baselines. We compared NVUM with several methods, including the CheXNet baseline [152], Ma et al.'s approach [131] based on a cross-attention network, the current SOTA for NIH on the official test set is the model by Hermoza et al. [68] that is a weakly supervised disease classifier that combines region proposal and saliency detection. We

[2] We include a detailed description based on [30] in the Appendix B.

Table 5.1: Class-level and mean testing AUC on OPI [37] and PDC [15] for the experiment based on training on NIH [192]. Best results for OPI/PDC are in **bold**/underlined.

Models	ChestXNet [152]		Hermoza et al. [68]		Ma et al. [131]		DivideMix [106]		Ours	
Datasets	OPI	PDC	OPI	PDC	OPI	PDC	OPI	PDC	OPI	PDC
Atelectasis	86.97	84.99	86.85	83.59	84.83	79.88	70.98	73.48	**88.16**	85.66
Cardiomegaly	89.89	92.50	89.49	91.25	90.87	91.72	74.74	81.63	**92.57**	92.94
Effusion	94.38	96.38	95.05	96.27	94.37	96.29	84.49	97.75	**95.64**	96.56
Infiltration	76.72	70.18	**77.48**	64.61	71.88	73.78	84.03	81.61	72.48	72.51
Mass	53.65	75.21	95.72	86.93	87.47	85.81	71.31	77.41	**97.06**	85.93
Nodule	86.34	75.39	82.68	75.99	69.71	68.14	57.45	63.89	**88.79**	75.56
Pneumonia	**91.44**	76.20	88.15	75.73	84.79	76.49	64.65	72.32	90.90	82.22
Pneumothorax	80.48	79.63	75.34	74.55	82.21	79.73	71.56	75.46	**85.78**	79.50
Edema	83.73	98.07	85.31	97.78	82.75	96.41	80.71	91.81	**86.56**	95.70
Emphysema	82.37	79.10	83.26	79.81	79.38	75.11	54.81	59.91	**83.70**	79.38
Fibrosis	90.53	96.13	86.26	96.46	83.17	93.20	76.98	84.71	**91.67**	98.40
Pleural Thickening	81.58	72.29	77.99	69.95	77.59	67.87	63.98	58.25	**84.82**	74.80
Hernia	89.82	86.72	93.90	89.29	87.37	86.87	66.34	72.11	**94.28**	93.02
Mean AUC	83.69	83.29	86.01	83.25	82.80	82.41	70.92	76.18	**88.65**	85.55

also show results from DivideMix [106], which uses a noisy-label learning algorithm based on small loss sample selection and semi-supervised learning. DivideMix has the SOTA results in many noisy-label learning benchmarks. All methods are implemented using the same DenseNet121 [72] backbone.

Quantitative Comparison Table 5.1 shows the class-level AUC result for training on NIH and testing on OPI and PDC. Our approach achieves the SOTA results on both clean test sets, consistently outperforming the baselines [68, 131, 152], achieving 2% mean AUC improvement on both test sets. Compared with the current SOTA noisy-label learning DivideMix [106], our method outperforms it by 18% on OPI and 9% on PDC. This shows that for noisy multi-label imbalanced MIA datasets, noisy multi-class balanced approaches based on small-loss selection is insufficient because they do not take into account the multi-label and imbalanced characteristics of the datasets. Table 5.2 shows class-level AUC results for training on CXP and testing on OPI and PDC. Similarly to the NIH results on Table 5.1, our approach achieves the best AUC results on both test sets with at least 3% improvement on OPI and 3% on PDC. In addition, DivideMix [106] shows similar results compared with NIH. Hence, SOTA performance on both noisy training sets suggests that our method is robust to different noisy multi-label imbalanced training sets.

Additional benchmark. Using the recently proposed noisy label benchmark by Xue et al. [217], we further test our approach against the SOTA in the field. The benchmark uses a subset of the official NIH test set [133], with 1,962 CXR images manually re-labelled by at least three radiologists per image. For the results, we follow [217] and consider the AUC results only for Pneumothorax (Pneu) and average of Mass and Nodule (M/N). We use the same hyperparameters as above. The results in Tab. 5.3

Table 5.2: Class-level and mean testing AUC on OPI [37] and PDC [15] for the experiment based on training on CXP [76].Best results for OPI/PDC are in **bold**/underlined.

Methods	CheXNet [152]		Hermoza et al. [68]		Ma et al. [131]		DivideMix[106]		Ours	
Datasets	OPI	PDC	OPI	PDC	OPI	PDC	OPI	PDC	OPI	PDC
Cardiomegaly	84.00	80.00	87.01	87.20	82.83	85.89	71.14	66.51	**88.86**	88.48
Edema	88.16	98.80	87.92	98.72	86.46	97.47	75.36	95.51	**88.63**	99.60
Pneumonia	**65.82**	58.96	65.56	53.42	61.88	54.83	57.65	40.53	64.90	67.89
Atelectasis	77.70	72.23	78.40	75.33	80.13	72.87	73.65	64.12	**80.81**	75.03
Pneumothorax	77.35	84.75	62.09	78.65	51.08	71.57	68.75	54.05	**82.18**	83.32
Effusion	85.81	91.84	87.00	93.44	**88.43**	92.92	78.60	79.89	83.54	89.74
Fracture	57.64	60.26	57.47	53.77	59.92	60.44	**60.35**	59.43	57.02	62.67
Mean AUC	76.64	78.12	75.06	77.29	72.96	76.57	69.36	65.72	**77.99**	80.96

Table 5.3: Pneumothorax and Mass/Nodule AUC using the manually labelled clean test from [133]. Baseline results obtained from [217]. Best results are in **bold**.

	BCE	F-correction [148]	MentorNet [82]	Decoupling [134]	Co-teaching [60]	ELR [119]	Xue et al. [217]	Ours
Pneu	87.0	80.8	86.6	80.1	87.3	87.1	**89.1**	**88.9**
M/N	84.3	84.8	83.7	84.3	82.0	83.2	84.6	**85.5**

shows that our method outperforms most noisy label methods and achieves comparable performance to [217] on Pneumothorax (88.9 vs 89.1) and better performance on Mass/Nodule (85.5 vs 84.6). However, it is important to mention that differently from [217] that uses two models, we use only one model, so our method requires significantly less training time and computation resources. Furthermore, the clean test set from [133] is much smaller than OPI and PDC with only two classes available, so we consider results in Tab. 5.1 and 5.2 more reliable than Tab. 5.3

5.3.2 Ablation Study

Different components of NVUM with π. We first study in Fig. 5.2 (left) how results are affected by the prior added on different components of NVUM. We run each experiment three times and show mean and standard deviation of AUC results. By adding the class prior π to ℓ_{BCE} [135], we replace the BCE term in (5.1) with $\ell_{BCE}(\tilde{\mathbf{y}}^i, \sigma(f_\theta(\mathbf{x}^i + \log \pi)))$. We can also add the class prior π to ℓ_{REG} by replacing the regularization term in (5.1) with $\ell_{REG}(\mathbf{t}^i, \sigma(f_\theta(\mathbf{x} + \log \pi)))$. We observe a 2% improvement for OPI and PDC for both modifications compared to ℓ_{BCE} baseline, demonstrating that it is important to handle imbalanced learning in MIA problems. Furthermore, we combine two modifications together and achieve additional 1% improvement. However, instead of directly working on the loss functions, as suggested in [135], we work on the memory module given that it also enforces the early learning phenomenon, addressing the combined noisy multi-label imbalanced learning problem.
Different β. We also study different values for β in (5.3). First, we test a volatile

ℓ_{BCE}	π	ℓ_{REG}	π	OPI	PDC
✓				82.91±0.78	82.27±1.02
✓	✓			85.80±0.04	84.35±0.35
✓		✓		85.24±0.70	84.39±0.21
✓	✓	✓		85.36±0.11	83.04±0.79
✓	✓	✓	✓	86.68±0.16	85.02±0.18
NVUM				**88.17±0.48**	**85.49±0.06**

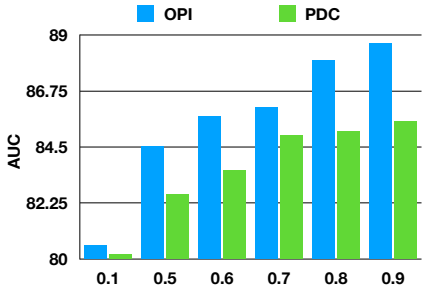


Figure 5.2: (Left) Mean AUC results of training on NIH using the class prior distribution π applied to different components of NVUM. (Right) Mean AUC results on OPI (blue) and PDC (green) of training on NIH with different β values for the NVUM memory update in (5.3).

memory update with $\beta = 0.1$, which shows a significantly worse performance because the model is overfitting the noisy multi-label of the training set. This indicates traditional volatile memory [63] cannot handle noisy label learning. Second, the non-volatile memory update with $\beta \in \{0.5, ..., 0.9\}$ shows a performance that improves consistently with larger β. Hence, we use $\beta = 0.9$ as our default setup.

5.4 Conclusions and Future Work

In this work, we argue that the MIA problem is a problem of *noisy multi-label* and *imbalanced learning*. We presented the Non-volatile Unbiased Memory (NVUM) training module, which stores a non-volatile running average of model logits to make the learning robust to noisy multi-label datasets. Furthermore, The NVUM takes into account the class prior distribution when updating the memory module to make the learning robust to imbalanced learning. We conducted experiments on proposed new benchmark and recent benchmark [217] and achieved SOTA results. Ablation study shows the importance of carefully accounting for imbalanced and noisy multi-label learning. For the future work, we will explore an precise estimation of class prior π during the training for accurate unbiasing.

Statement of Authorship

Title of Paper	BoMD: Bag of Multi-label Local Descriptors for Noisy Chest X-ray Classification
Publication Status	☐ Published ☑ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Published at International Conference on Computer Vision (ICCV), 2022

Principal Author

Name of Principal Author (Candidate)	Fengbei Liu		
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper and revision		
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	09/14/2023

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 in 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 and revision		
Signature		Date	09/15/2023

Name of Co-Author	Hu Wang
Contribution to the Paper	Discussion and commented on the revision
Signature	Date

Name of Co-Author	Yu Tian
Contribution to the Paper	Discussion and commented on the revision
Signature	Date

Name of Co-Author	Chong Wang
Contribution to the Paper	Discussion and commented on the revision
Signature	Date

Name of Co-Author	Yuyuan Liu
Contribution to the Paper	Discussion and commented on the revision
Signature	Date

Name of Co-Author	Gustavo Carneiro
Contribution to the Paper	Discussion and wrote the revision
Signature	Date 22/09/2023

Chapter 6

BoMD: Bag of Multi-label Descriptors for Noisy Chest X-ray Classification

Abstract

Deep learning methods have shown outstanding classification accuracy in medical imaging problems, which is largely attributed to the availability of large-scale datasets manually annotated with clean labels. However, given the high cost of such manual annotation, new medical imaging classification problems may need to rely on machine-generated noisy labels extracted from radiology reports. Indeed, many Chest X-Ray (CXR) classifiers have been modelled from datasets with noisy labels, but their training procedure is in general not robust to noisy-label samples, leading to sub-optimal models. Furthermore, CXR datasets are mostly multi-label, so current multi-class noisy-label learning methods cannot be easily adapted. In this paper, we propose a new method designed for noisy multi-label CXR learning, which detects and smoothly relabels noisy samples from the dataset to be used in the training of common multi-label classifiers. The proposed method optimises a bag of multi-label descriptors (BoMD) to promote their similarity with the semantic descriptors produced by language models from multi-label image annotations. Our experiments on noisy multi-label training sets and clean testing sets show that our model has state-of-the-art accuracy and robustness in many CXR multi-label classification benchmarks, including a new benchmark that we propose to systematically assess noisy multi-label methods. Code is available at https://github.com/cyh-0/BoMD.

6.1 Introduction

The promising results produced by deep neural networks (DNN) in medical image analysis (MIA) problems [113] is attributed to the availability of large-scale datasets with accurately annotated images. Given the high cost of acquiring such datasets, the field

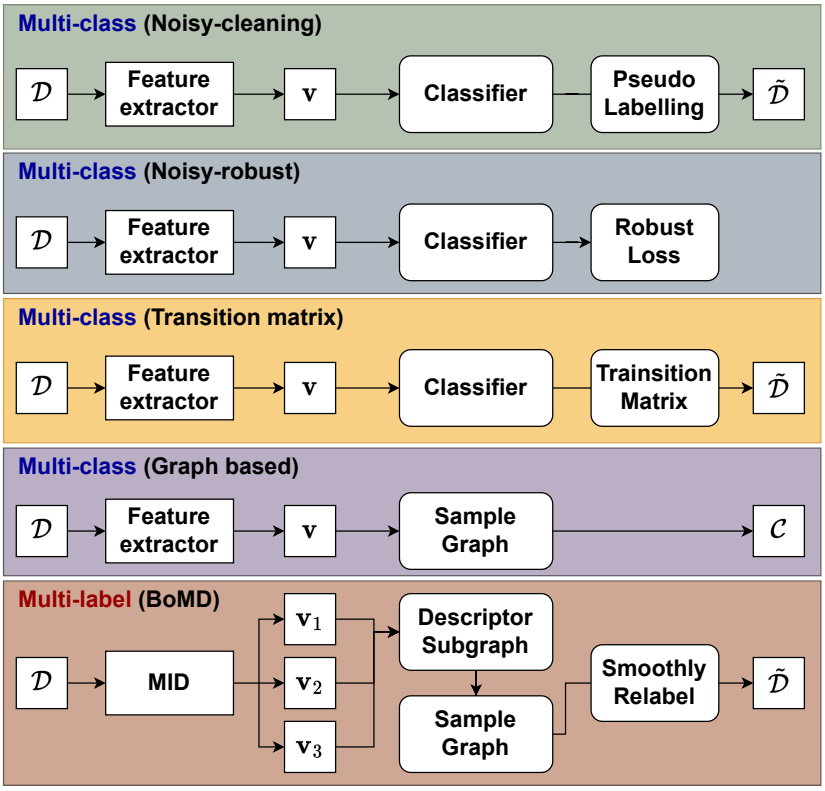


Figure 6.1: Comparison of multi-class LNL methods [5, 77, 106, 119] and our noisy multi-label approach, BoMD, where the feature extractor returns a single descriptor $\mathbf{v}$ per image, $\mathcal{D}$ is the noisy training set, $\mathcal{C}$ is the clean set, and $\tilde{\mathcal{D}}$ is the re-labelled training set. BoMD has two components: 1) learning of a bag of multi-label image descriptors (MID) $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ to represent the image, and 2) smooth re-labelling of images driven by a graph structure based on the fine-grained relationships between MID descriptors.

is considering more affordable automatic annotation processes by Natural Language Processing (NLP) approaches that extract multiple labels (each label representing a disease) from radiology reports [75, 193]. However, mistakes made by NLP combined with uncertain radiological findings can introduce label noise [143, 144], as can be found in NLP-annotated Chest X-ray (CXR) datasets [75, 193] whose noisy multi-labels can mislead supervised training processes. Nevertheless, even without addressing this noisy multi-label problem, current CXR multi-label classifiers [8, 68, 131, 152] show promising results. Although these methods show encouraging multi-label classification accuracy, there is still potential for improvement that can be realised by properly addressing the noisy multi-label learning problem present in CXR datasets [75, 193].

Current learning with noisy label (LNL) approaches focus on leveraging the hidden clean-label information to assist the training of DNNs (see Fig. 6.1). This can be achieved with techniques that clean the label noise [91, 106], robustify loss functions [77, 114, 119], estimate label transition matrices [5, 54, 209, 221], smooth training labels [128, 200, 227], and use graphs to explore the latent structure of data. [77, 204, 207]. These methods have been designed for noisy multi-class problems and do not easily extend to *noisy multi-label* learning, which is challenging given the potential multiple label mistakes for each training sample. In addition, a key characteristic of multi-label classification is the inherent positive-negative imbalance [154] issue. Such an issue may cause sample-selection based methods (e.g., DivideMix [106]) to select an extremely imbalanced clean set, where the majority of identified clean samples belong to the 'No Findings' class. Additionally, it could impede the accurate estimation of the posterior probabilities for the noisy or intermediate classes [108]." To the best of our knowledge, the state-of-the-art (SOTA) approach that handles noisy multi-label learning is NVUM [114], which is based on an extension of early learning regularisation (ELR) [119]. NVUM shows promising results, but it is challenged by the different early convergence patterns of multiple labels, which can lead to poor performance for particular label noise conditions, as shown in our experiments. Additionally, NVUM is only evaluated on real-world CXR datasets [75, 193] without any systematic assessment of robustness to varying label noise conditions, preventing a more complete understanding of its functionality.

In this paper, we propose a new solution specifically designed for the noisy multi-label problem by answering the following question: ***can the detection and correction of noisy multi-labelled samples be facilitated by leveraging the semantic information of training labels?*** available from language models [74, 104, 149]? This question is motivated by the successful exploration of language models in computer vision [8, 32, 70, 232], with methods that leverage semantic information to influence the training of visual descriptors; an idea that has not been explored in noisy multi-label classification. To answer this question, we introduce the 2-stage Bag of Multi-label Descriptors (BoMD) method (see Fig. 6.1) to smoothly re-label noisy multi-label image

datasets that can then be used for training common multi-label classifiers. The first stage trains a feature extractor to produce a bag of multi-label image descriptors by promoting their similarity with the semantic embeddings from language models. For the second stage, we introduce a novel graph structure, where each image is represented by a sub-graph built from the multi-label image descriptors, learned in the first stage, to smoothly re-label the noisy multi-label images. Compared with graphs built directly from a single descriptor per image [77], our graph structure with the multi-label image descriptors has the potential to capture more fine-grained image relationships, which is crucial to deal with multi-label annotation. We also propose a new benchmark to systematically assess noisy multi-label methods. In summary, our **contributions** are:

1. A novel 2-stage learning method to smoothly re-label noisy multi-label datasets of CXR images that can then be used for training a common multi-label classifier;

2. A new bag of multi-label image descriptors learning method that leverages the semantic information available from language models to represent multi-label images and to detect noisy samples;

3. A new graph structure to smoothly re-label noisy multi-label images, with each image being represented by a sub-graph of the learned multi-label image descriptors that can capture fine-grained image relationships;

4. The first systematic evaluation of noisy multi-label methods that combine the PadChest [16] and Chest X-ray 14 [193] datasets.

We show the effectiveness of our BoMD on a benchmark that consists of training with two noisy multi-label CXR datasets and testing on three clean multi-label CXR datasets [114]. Results show that our approach has more accurate classifications than previous multi-label classifiers developed for CXR datasets and noisy-label classifiers. Results on our proposed benchmark show that BoMD is generally more accurate and robust than competing methods under our systematic evaluation.

6.2 Related Works

6.2.1 CXR multi-label classification

Recently, we have seen many CXR multi-label classifiers being proposed, such as the CXR pneumonia detector [152]. Ma et al. [131] introduce a new cross-attention network to extract meaningful representations. Hermoza et al. [68] propose a weakly-supervised method to diagnose and localise diseases. Although these methods show promising results, there is still potential for improvement that can be realised by addressing the noisy multi-label learning of CXR datasets [75, 193].

6.2.2 Learning with Noisy Labels

Noise-cleaning methods focus on detecting noisy samples. For instance, Han et al. [60] rely on the small-loss trick (i.e., clean samples have small losses) to co-teach two models. Huang et al. [73] detect noisy samples that have unstable prediction by switching learning rates. Bahri et al. [6] discard samples whose labels disagree with a KNN classifier prediction. Noise-cleaning methods can be combined with semi-supervised learning [12] to perform both the detection and correction of corrupted data. For example, DivideMix [106] removes the labels of samples classified as noisy and runs a semi-supervised learning method [12, 125]. FINE [91] proposes a robust method to detect noisy samples by verifying the alignment of image features and class-representative eigenvectors. Noise-cleaning methods generally employ two divergent networks to reduce confirmation bias [106, 124], which substantially increases computational complexity. Additionally, it is unclear if these methods can handle noisy multi-label problems since they do not in general capture fine-grained image relationships.

Noise-robust methods rely on robust loss functions to balance the overfitting effects caused by label noise in the training process. Early papers, such as [196], explore the symmetric property of cross-entropy (CE) loss for noise-robust learning. Zhang et al. [235] propose the combination of Mean-Absolute-Error (MAE) and cross-entropy (CE) loss to achieve a good balance between convergence and generalisation. Ma et al. [132] show that any loss function can be robust to label noise by applying a simple normalization term. Recently, [44] proposes a noise-robust Jensen-Shannon divergence (JSD) loss based on a soft transition between MAE and CE losses. Even though these methods can reduce overfitting effects, they also tend to under-fit the training data. This issue has been partially addressed by the early learning regularisation (ELR) [119] that proposes a regularization term which restricts the gradient from samples with corrupt labels. The non-volatile unbiased memory (NVUM) [114] extends ELR to noisy multi-label problems. Although promising, ELR and NVUM are challenged by the different early convergence patterns of multiple labels, which can lead to poor performance for particular label noise conditions, as shown in the experiments.

Transition matrix methods estimate the transition probability between clean and noisy labels. Goldberger et al. [54] propose a noise adaptation layer to estimate label transition. Yao et al. [221] estimate the transition matrix using an intermediate class and a factorised matrix. Xia et al. [209] estimate part-dependent transition matrix for complex noise conditions. Bae et al. [5] proposed a noisy prediction calibration method based on a transition matrix to reduce the gap between noisy prediction and clean label based on a KNN prediction.

Label-smoothing methods rely on modifying [137, 174] the sample-wise label distribution [128]. Zhang et al. [227] propose online label smoothing (OLS) which generates soft labels by considering the relationships among multiple labels. Wei et al. [200] argue

that the advantage of label smoothing vanishes under a high label noise regime since the label smoothing tends to over-smooth the estimated label classification, so they propose the generalised label smoothing (GLS) [200] which uses negative smoothing values for higher noise rates. In general, label smoothing methods can underfit the training data since they tend to abandon the optimisation of hard clean-label samples [137].

Graph-based methods leverage the robustness of feature representations to discriminate between clean and noisy samples and regularise the training process. Wu et al. [204] explore the topological property of the data in the feature space to perform noise-cleaning by assuming that clean data are clustered together in this feature space, while the corrupted data are isolated. Wu et al. [207] investigate the geometric structure of the data to model predictive confidence and filter out noisy samples. Iscen et al. [77] introduce a regularisation term that forces samples to have similar predictions to their neighbors. These graph-based methods have been designed for single-label classification, so they cannot be easily adapted to multi-label datasets. Also, building a graph with multi-label data is also an issue for these methods.

Multi-label Noisy label methods have received increasing attention in recent years due to the natural differences with respect to multi-class problems. Instead of the sample-wise noise found in multi-class problems, each label per sample can be corrupted in the multi-label scenario which can be problematic for selecting and correcting the label noise. In addition, the class imbalance [114] and the semantic divergence may also exacerbate the overfitting issue towards the majority classes. Zhao et al. [236] leverage label dependencies to handle noisy labels and use word embeddings to perform context-based regularization to avoid overfitting. Li et al. [108] consider the correlation between labels (i.e., "fish" and "water" have a stronger correlation when comparing with "fish" and "sky") to estimate the transition matrix. Xie et al. [212] mitigate the negative impact of label noise by estimating the confidence for credible labels from the candidate label set. Different from previous methods, we consider using the label semantic information and label smoothing techniques to capture more fine-grained image relationships and prevent the classifier from being overconfident on any of the noisy labels.

6.2.3 Bag of Words

The Bag of Words (BoW) method [62, 167, 168] is a traditional information retrieval technique, denoted by the representation of documents with a histogram of unordered words. In computer vision, BoW [33, 167] represents images with a histogram of unordered local visual descriptors, learned from the training images in an unsupervised manner. We adopt the BoW concept, but instead of extracting local visual descriptors (e.g., SIFT [127]), we train a DNN to represent each image with a bag of global visual descriptors.

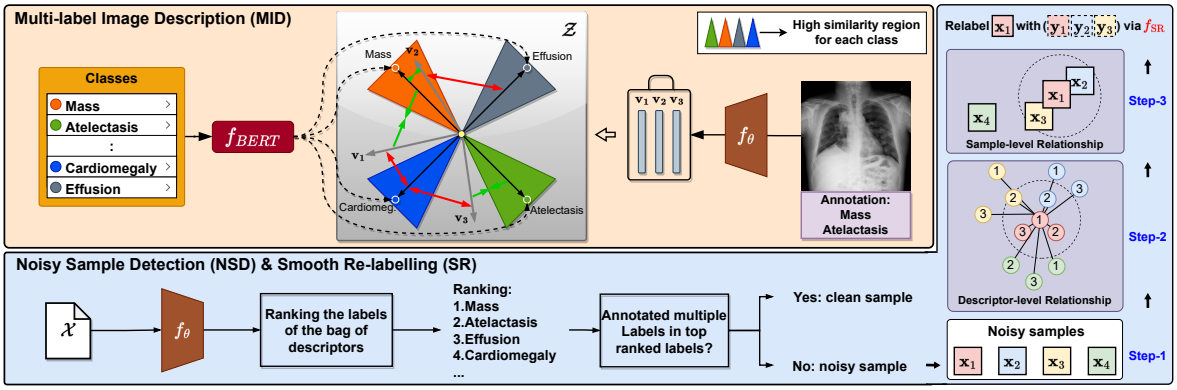


Figure 6.2: The **multi-label image description (MID)** module estimates a bag of visual descriptors $\{\mathbf{v}_m\}_{m=1}^{M}$ for each image, which is learned using the semantic information provided by BERT [46, 83, 149]. BERT represents each class with a descriptor (circle) in the semantic space $\mathcal{Z}$, where the triangular regions indicate high similarity regions of each corresponding class. Please note that the green arrows represent the pulling of image descriptors towards one of the corresponding relevant embeddings in the rank loss of Eq.2. Conversely, the red arrows push the image descriptors away from all irrelevant embeddings and the rectangular regions indicate the high similarity area w.r.t to each class's word embeddings. The **noisy sample detection (NSD)** module leverages the consistency between label ranking prediction from MID and the original annotation to detect the noisy samples, which are then **smoothly re-labelled (SR)** with $f_{\text{SR}}(.)$ in (6.6).

6.3 Method

We assume the availability of a noisy multi-label training set denoted by $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^{H \times W \times R}$ represents an image of size $H \times W$ and R colour channels[1], and $\mathbf{y}_i \in \mathcal{Y} = \{0, 1\}^{|\mathcal{Y}|}$ denotes the multi-label annotation. The testing set is similarly defined.

6.3.1 Bag of Multi-label Descriptors (BoMD)

Our method, described in Alg. 2, is inspired by the observation that a noisy-labelled sample tends to be an outlier surrounded by clean-labelled samples in the feature space [204]. Hence, the label of each sample should be consistent with the labels of the neighboring samples. This motivated us to develop an approach to re-label a noisy multi-label image with the estimated label distribution from its neighbourhood. The proposed BoMD has two stages (Fig. 6.1): 1) image description learning that transforms a training image into a bag of visual descriptors that lie in the semantic space $\mathcal{Z} \subset \mathbb{R}^Z$ populated by word embeddings computed from image labels [149] 2) graph construction to smoothly re-label noisy multi-label images, where each image is represented by a sub-graph built from the learned bag of visual descriptors, which can capture fine-grained image relationships. This smoothly re-labelled dataset is then used for training a multi-label classifier.

6.3.2 Multi-label Image Description (MID)

Motivated by BoW, our MID (Fig. 6.2) represents an image by associating its multiple labels to a bag of global visual descriptors. MID projects the image into BERT's semantic space using a set of visual descriptors that are optimised to promote their similarity with the semantic descriptors produced by BERT models from the multi-label annotation of the image. The MID of image $\mathbf{x}$ are extracted with $\mathcal{V} = f_\theta(\mathbf{x})$, where $\mathcal{V} = \{\mathbf{v}(m)\}_{m=1}^{M}$ denotes the M[2] visual descriptors in BERT's semantic space, i.e., $\mathbf{v}(m) \in \mathcal{Z}$. The BERT language models (e.g., BlueBERT [149], medical language model pre-trained on PubMed abstracts [46] and clinical notes [83]) produce semantic descriptors in the form of word embeddings with $\mathbf{w}(c) = f_{BERT}^{\mathcal{Y}}(c)$ for $c \in \{1, ..., |\mathcal{Y}|\}$, forming $\mathcal{W} = \{\mathbf{w}(c)\}_{c=1}^{|\mathcal{Y}|}$, where $\mathbf{w}(c) \in \mathcal{Z}$, with $\mathcal{Z}$ being the same space as for $f_\theta(.)$. More specifically, MID is trained with:

$$\theta^* = \arg\min_\theta \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}} \omega_i \ell_{mid}(\mathbf{x}_i, \mathbf{y}_i, \theta) + \beta \ell_{reg}(\mathbf{x}_i, \theta) \tag{6.1}$$

[1] We consider each image to have 3 colour channels (R=3) and dimensions of $H = W = 512$ for NIH, and $H = W = 224$ for CXP.

[2] We empirically set $M = 3$ according the ablation study of the hyper-parameters in Supp. Material.

where $\omega_i = \left(\left(\sum_{c=1}^{|\mathcal{C}|} \mathbb{I}(\mathbf{y}_i(c)=1)\right) \div \left(\sum_{c=1}^{|\mathcal{C}|} \mathbb{I}(\mathbf{y}_i(c)=0)\right)\right)$ is a normalisation that controls the ranking weight based on the number of positive and negative labels ($\mathbb{I}(.)$ represents an indicator function) [14, 232], and the hyper-parameter β weights the regulariser. Also in (6.1) we have:

$$\begin{aligned} &\ell_{mid}(\mathbf{x}_i, \mathbf{y}_i, \theta) = \\ &\sum_{\substack{p=1,\\ \mathbf{y}_i(p)=1}}^{|\mathcal{Y}|} \sum_{\substack{n=1,\\ \mathbf{y}_i(n)=0}}^{|\mathcal{Y}|} \log\left(1+\exp\left(\ell_{rank}(\mathcal{V}_i, \mathcal{W}, p, n)\right)\right), \text{ where} \\ &\ell_{rank}(\mathcal{V}_i, \mathcal{W}, p, n) = \left(\max_{\mathbf{v}\in\mathcal{V}_i}(\langle \mathbf{v}, \mathbf{w}(n)\rangle) - \max_{\mathbf{v}\in\mathcal{V}_i}(\langle \mathbf{v}, \mathbf{w}(p)\rangle)\right), \end{aligned} \tag{6.2}$$

with $\langle ., . \rangle$ representing the dot product operator, $p, n \in \{1, ..., |\mathcal{Y}|\}$ denoting the indices to the positive (i.e., $\mathbf{w}(p)$ where $\mathbf{y}_i(p) = 1$) and negative word embeddings (i.e., $\mathbf{w}(n)$ where $\mathbf{y}_i(n) = 0$), respectively, $\mathcal{V}_i = f_\theta(\mathbf{x}_i)$, and

$$\ell_{reg}(\mathbf{x}_i, \theta) = \sum_{\mathbf{v}(m)\in\mathcal{V}_i} \frac{(\mathbf{v}(m) - \bar{\mathbf{v}}(m))^\top (\mathbf{v}(m) - \bar{\mathbf{v}}(m))}{Z-1} \tag{6.3}$$

being a regulariser to reduce descriptor variance, where Z is the number of dimensions of $\mathcal{Z}$, and $\bar{\mathbf{v}}(m)$ denoting the MID mean in $\mathcal{V}_i$. The $\ell_{mid}(.)$ in (6.2), inspired by previous multi-label learning methods [232], forces the dot product $\langle \mathbf{v}, \mathbf{w}(p)\rangle$ to rank higher than the $\langle \mathbf{v}, \mathbf{w}(n)\rangle$, which means that the visual descriptors will be more similar to positive label embeddings than the negative label ones. Intuitively, this loss encourages semantically similar image descriptors to cluster around their related semantic descriptors, which will benefit our graph-based smooth re-labelling.

6.3.3 Graph Construction and Smooth Re-labelling

Considering that the learned visual descriptors are likely to be closer to clean labels in the semantic space, we formulate the detection of noisy training samples by first ranking (in descending order of similarity) the labels for image $\mathbf{x}_i$ (according to the inner product with the word embeddings from the labels), as follows:

$$\mathbf{r}_i = \arg\operatorname{sort}_{c\in\{1,...,|\mathcal{Y}|\}} \left[\max_{\mathbf{v}\in\mathcal{V}_i}(\langle \mathbf{v}, \mathbf{w}(c)\rangle)\right], \tag{6.4}$$

where $\mathbf{r}_i(1) \in \{1, ..., |\mathcal{Y}|\}$ is the highest ranked label and $\mathbf{r}_i(|\mathcal{Y}|) \in \{1, ..., |\mathcal{Y}|\}$ is the lowest ranked label. Then, clean samples are the ones where $\mathbf{r}_i(p) < \mathbf{r}_i(n)$ for all positive labels $p \in \mathcal{P}_i$ and negative labels $n \in \mathcal{N}_i$, with $\mathcal{P}_i = \{c|\mathbf{y}_i(c) = 1\}$ and $\mathcal{N}_i = \{c|\mathbf{y}_i(c) = 0\}$; otherwise, the sample is classified as noisy.

The second stage of BoMD re-labels noisy samples using the sample graph built with the MID visual descriptors. The graph is constructed by representing each training image $\mathbf{x}_i$ with M nodes $\{\mathbf{v}_i(m)\}_{m=1}^{M}$ from $\mathcal{V}_i = f_\theta(\mathbf{x}_i)$, where the edge weight between the m^{th} descriptor of the i^{th} image and the n^{th} descriptor of the j^{th} image is defined by $e(\mathbf{v}_i(m), \mathbf{v}_j(n)) = 1/\|\mathbf{v}_i(m) - \mathbf{v}_j(n)\|_2$. This means that the graph has the set of nodes denoted by $\{\mathbf{v}_i(m)\}_{i=1,m=1}^{|\mathcal{D}|,M}$ and edges $\{e(\mathbf{v}_i(m), \mathbf{v}_j(n))\}_{i,j=1,m,n=1}^{|\mathcal{D}|,M}$. The re-labelling is based on finding the K nearest neighbouring training images to image i by using the graph nodes, with:

$$\mathcal{K}(\mathcal{V}_i) = \text{top}K_{j\in\{1,\ldots,|\mathcal{D}|\},m,n\in\{1,\ldots,M\}}(e(\mathbf{v}_i(m), \mathbf{v}_j(n))), \tag{6.5}$$

where $\mathcal{K}(\mathcal{V}_i)$ contains the unique image indices from the K nodes with the largest edge weights. Next, for all samples identified as noisy, we update their labels with $\tilde{\mathbf{y}}_i = f_{\text{SR}}(\mathbf{y}_i, \bar{\mathbf{y}}_i)$, which is defined as

$$f_{\text{SR}}(\mathbf{y}_i, \bar{\mathbf{y}}_i) = (1-\lambda)\cdot\mathbf{y}_i + \lambda\cdot\left(\gamma\cdot\mathbf{1}_{|\mathcal{Y}|} + (1-\gamma)\cdot\bar{\mathbf{y}}_i\right)\odot\mathbf{m}, \tag{6.6}$$

where $\lambda \in [0,1], \gamma \in [0, 0.5]$, $\bar{\mathbf{y}}_i = \frac{1}{K}\sum_{j\in\mathcal{K}(\mathcal{V}_i)}\mathbf{y}_j$, $\mathbf{1}_{|\mathcal{Y}|}$ denotes a vector with ones of size $|\mathcal{Y}|$ (uniform distribution) to prevent the re-labelling from being overconfident on any of the labels, $\odot$ is the element-wise vector multiplication, and $\mathbf{m} = \mathbb{I}((\mathbf{y}_i + \bar{\mathbf{y}}_i) > 0)$ is a binary mask to filter out high confident negative labels (with $\mathbb{I}(.)$ being the indicator function) to mitigate the over-smoothing issue.

6.3.4 Training and Testing

We build a new training set $\tilde{\mathcal{D}} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i)|(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}\}_{i=1}^{|\mathcal{D}|}$, where $\tilde{\mathbf{y}}_i = \mathbf{y}_i$ if sample $(\mathbf{x}_i, \mathbf{y}_i)$ is clean from Eq. (6.4), or computed from Eq. (6.6) if sample is noisy}. Then, we train a regular classifier $f_\phi : \mathcal{X} \to [0,1]^{|\mathcal{Y}|}$ by minimizing a BCE loss on $\tilde{\mathcal{D}}$. Testing is based on applying the trained $f_\phi(.)$ to test images.

6.4 Experiments

Our experiments are based on the following datasets. **Noisy Training Sets.** The ***NIH Chest X-ray14*** **(NIH)** [193] contains 112,120 frontal-view CXR images from 30,805 patients, where each image has between 0 and 14 annotated pathologies and the training set contains 86,524 images with a maximum of 9 labels per image. The ***CheXpert (CXP)*** [75] has 224,316 frontal-view CXR images from 65,240 patients labelled with 14 common chest radiographic observations, where the training set contains 170,958 images with a maximum of 8 labels per image. The labels of these two datasets are obtained from an NLP algorithm, which forms noisy multi-label annotations [143].

Algorithm 2 BoMD

require: Training set $\mathcal{D}$, and BERT model embeddings $\mathbf{w}(c) = f^{\mathcal{Y}}_{BERT}(c)$ for $c \in \{1, ..., |\mathcal{Y}|\}$
Build MID
Train $f_\theta(.)$ with (6.1) using $f^{\mathcal{Y}}_{BERT}(.)$ and $\mathcal{D}$
Create re-labeled set $\tilde{\mathcal{D}} = \emptyset$ and noisy set $\mathcal{B} = \emptyset$
# Detect noisy samples
for $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}$ **do**
    Compute label rank $\mathbf{r}_i$ from (6.4)
    If $\mathbf{r}_i(p) > \mathbf{r}_i(n)$ for all $p \in \mathcal{P}_i$ and all $n \in \mathcal{N}_i$
    Then $\tilde{\mathcal{D}} \leftarrow \tilde{\mathcal{D}} \bigcup (\mathbf{x}_i, \mathbf{y}_i)$
    Else $\mathcal{B} \leftarrow \mathcal{B} \bigcup (\mathbf{x}_i, \mathbf{y}_i)$
end for
Build graph with nodes $\{\mathbf{v}_i(m)\}_{i=1,m=1}^{|\mathcal{D}|,M}$ and edges $\{e(\mathbf{v}_i(m), \mathbf{v}_j(n))\}_{i,j=1,m,n=1}^{|\mathcal{D}|,M}$ using $f_\theta(.)$ and $\mathcal{D}$
# Re-label noisy samples
for $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{B}$ **do**
    Re-label $\mathbf{x}_i$ with $\tilde{\mathbf{y}}_i$ from (6.6)
    $\tilde{\mathcal{D}} \leftarrow \tilde{\mathcal{D}} \bigcup (\mathbf{x}_i, \tilde{\mathbf{y}}_i)$
end for
# Train final classifier
Train $f_\phi : \mathcal{X} \rightarrow [0, 1]^{|\mathcal{Y}|}$ with BCE loss using $\tilde{\mathcal{D}}$

Clean Testing Sets. The ***OpenI*** [38] dataset contains 3,999 radiology reports and 7,470 frontal/lateral-view CXR images from the Indiana Network for Patient Care. We use all frontal-views images for evaluation, resulting in 3,818 images and 19 manually annotated diseases. We also use ***PadChest*** [16], which contains 160,861 images with 27 chest radiographic observations. PadChest has a mixture of machine and manually labelled images, but we only use the manually labelled frontal-view images (about 15.25%[3] of the images). Additionally, we follow previous works [114, 217] to evaluate the model on a re-organised subset of the official NIH test set [193], referred to as ***NIH-GOOGLE***, with 1,962 CXR images that focuses on two findings (pneumothorax and nodule/mass). Each image is manually re-labelled by at least three certified radiologists [133] with final label pooled from an adjudication process.
Systematic Noisy-label Assessment. We introduce the first systematic noisy multi-label assessment benchmark by combining the NIH [193] dataset and the clean test samples from Padchest [16] as a new noisy-label training set with the OpenI [38] as the clean testing set. We then apply symmetric label noise [5, 106] to the PadChest training subset **only**, where we flip the labels from present to absent, and vice-versa, based on two control variables, namely: 1) the proportion of noisy samples, and 2) the probability of switching a label. This dataset is referred to as ***NIHxPDC***.

6.4.1 Implementation Details

We resize the NIH [193] images to 512×512, and CXP [75] images to 224×224, where images are normalised using ImageNet [158] mean and standard deviation. We use random resize crop and random horizontal flipping as data augmentation. The BlueBERT word embeddings in $\mathcal{W}$ are L2 normalised. For the MID model $f_\theta(.)$, we use the ImageNet [158] pre-trained DenseNet121 [72], which is trained with Adam optimiser [93] using a learning rate of 0.0001 with cosine annealing decay [126], batch size of 16 and 20 epochs. We set the number of descriptors $M = 3$ and weight of regulariser $\beta = 0.3$. The descriptor graph is implemented with Faiss [84] for efficient search (the search process in Eq. (6.5) takes 5 seconds for all training samples from NIH). The classifier $f_\phi(.)$ uses another ImageNet pre-trained DenseNet121 [72], and then it is trained with Adam optimiser [93] using a learning rate of 0.05, batch size of 16 and 30 epochs. We empirically set the mixup coefficient λ and γ in Eq. (6.6) to 0.6 and 0.25, and use $K = 10$ in Eq. (6.5). The classification results are assessed with the mean of the class-wise area under the receiver operating characteristic curve (AUC) for all disease classes [68, 75, 116]. All experiments are implemented with Pytorch [147] and run on an NVIDIA RTX3090 GPU (24GB). Training takes 23h for NIH and 15h for CheXpert, and testing for a single image takes 13.41ms for NIH and 12.24ms for CheXpert.

[3]The 27% of PadChest's annotated images include the lateral-view CXRs.

Methods	Models	OpenI	PadChest
General	Hermoza et al [68]	85.54 ± 0.42	83.90 ± 0.57
	CAN [131]	84.26 ± 0.35	83.10 ± 0.25
Noise-cleaning	DivideMix [106]	72.76 ± 1.09	75.49 ± 0.21
	FINE [91]	63.67 ± 1.78	70.91 ± 0.20
Noise-robust	ELR [119]	86.62 ± 0.87	85.24 ± 0.11
	NVUM [114]	**88.17 ± 0.48**	**85.49 ± 0.06**
Transition matrix	NPC [5]	86.21 ± 0.07	83.88 ± 0.05
Graph-based	NCR [77]	85.06 ± 0.96	83.79 ± 0.48
Label smoothing	LS [128]	83.72 ± 1.29	80.93 ± 0.82
	OLS [227]	85.08 ± 0.31	83.51 ± 0.61
	GLS [200]	83.80 ± 0.34	81.56 ± 0.24
	BoMD	**89.57 ± 0.22**	**86.45 ± 0.08**

Table 6.1: Mean ± standard deviation AUC results for the testing sets from OpenI and PadChest, using models **trained on NIH [193]**. Best and the second best results are in **red**/**blue**.

6.4.2 Classification Results on Real-world Datasets

We first compare the performance of state-of-the-art (SOTA) methods with our BoMD in Table 6.1 and Table 6.2. We run each experiment three times and show mean and standard deviation of AUC results. Table 6.1 shows the testing AUC results of the training with NIH [193] and testing on OpenI [38] and PadChest [16]. Our model surpasses the second best method [114] by 1.4% and 0.96% on the two test sets with p-values 0.0018 and 10^{-14}, respectively (one-sided t-test). We also report testing performance on OpenI [38] and PadChest [16] using training on CXP [75] in Table 6.2. Our model surpasses the second best result [75] by 2.82% and 1.13% on the two test sets with p-values 0.0002 and 10^{-14}, respectively (one-sided t-test). We also notice that there is a large gap between all models' performance for certain classes. For example, in our model, the AUC results for *Pneumonia* classification when training on NIH are much better than when training on CXP, with a gap of 23.20% and 15.48% on OpenI [38] and PadChest [16], respectively, which may be due to CXP's smaller image size. The NIH-GOOGLE [114, 217] evaluation on classes pneumothorax and nodule/mass (obtained from the average classification scores for Mass and Nodule) is displayed in Table 6.3, which shows that our method outperforms the SOTA methods on both Pneumothorax (+0.6% compared to [217]) and Mass/Nodule (+2.4% compared [114]) classifications.

Methods	Models	OpenI	PadChest
General	Hermoza et al [68]	74.94 ± 0.50	77.24 ± 0.04
	CAN [131]	76.34 ± 0.49	78.92 ± 0.58
Noise-cleaning	DivideMix [106]	73.23 ± 0.60	74.21 ± 0.55
	FINE [91]	71.68 ± 0.54	73.83 ± 0.52
Noise-robust	ELR [119]	77.16 ± 0.79	79.97 ± 0.71
	NVUM [114]	**77.21 ± 0.81**	**80.62 ± 0.10**
Transition matrix	NPC [5]	75.32 ± 0.40	77.30 ± 0.03
Graph-based	NCR [77]	76.93 ± 0.38	79.36 ± 0.84
Label Smoothing	LS [128]	72.86 ± 0.23	75.34 ± 0.50
	OLS [227]	76.52 ± 0.83	77.72 ± 0.70
	GLS [200]	76.50 ± 0.26	78.80 ± 0.70
	BoMD	**80.03 ± 0.73**	**81.76 ± 0.40**

Table 6.2: Mean ± standard deviation AUC results for the testing sets from OpenI and PadChest, using models **trained on CXP [75]**. Best and the second best results are in **red**/**blue**.

	BCE	F-correction [148]	MentorNet [82]	Decoupling [134]	Co-teaching [60]	ELR [119]	Xue et al. [217]	NVUM [114]	BoMD
Pneu	87.0	80.8	86.6	80.1	87.3	87.1	**89.1**	88.9	**89.7**
M/N	84.3	84.8	83.7	84.3	82.0	83.2	84.6	**85.5**	**87.9**

Table 6.3: Pneumothorax and Mass/Nodule AUC of NIH-Google [133] for models trained on NIH [193]. Best and the second best results for OpenI and PadChest are in **red**/**blue**.

The per-finding results are reported in the *Appendix C*.

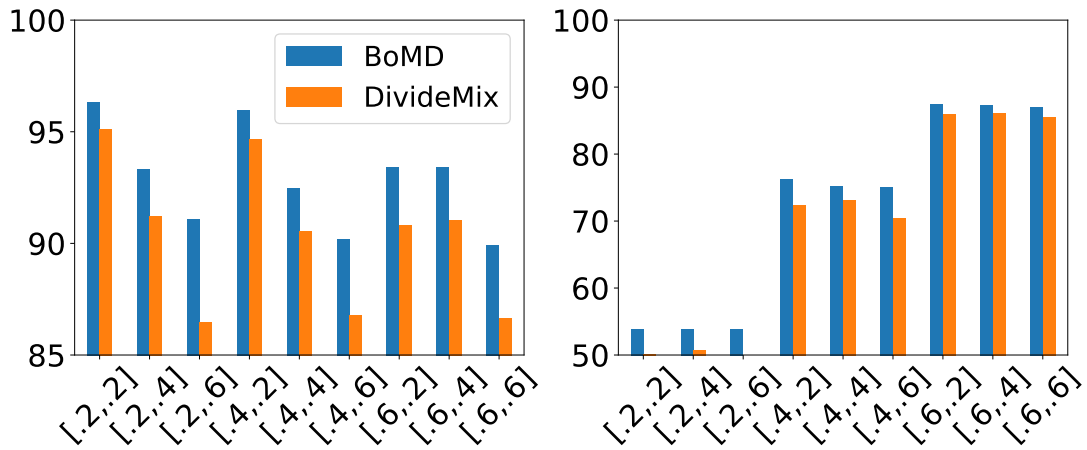
6.4.3 Systematic Noisy-label Benchmark

In our systematic noisy-label benchmark NIHxPDC, we compare our BoMD with the SOTA method NVUM [114] and a baseline model trained with BCE loss. Recall that this benchmark relies on two control variables: a) percentage of noisy samples p_s, and b) probability of switching a label p_l, where $p_s, p_l \in \{0\%, 20\%, 40\%, 60\%\}$. We show the mean AUC classification over the 14 classes on clean OpenI in Tab. 6.4. Notice that our BoMD has better results than NVUM and BCE for the majority of the cases, with a 3% to 5% improvement compared with NVUM and BCE when $p_s = 20\%$, 1% to 4% improvement for $p_s = 40\%$, and 1% to 5% improvement for $p_s = 60\%, p_l = 20\%, 40\%,$

p_s	0%	20%			40%			60%		
p_l	0%	20%	40%	60%	20%	40%	60%	20%	40%	60%
BCE	85.65	83.99	81.42	79.63	80.99	78.51	75.79	77.14	75.35	72.39
NVUM	87.89	85.34	82.83	81.35	82.52	80.64	78.66	78.49	77.19	**76.91**
BoMD	**89.76**	**88.00**	**86.26**	**84.55**	**84.47**	**81.86**	**78.68**	**82.23**	**78.15**	74.11

Table 6.4: Mean testing AUC results for the 13 OpenI classes with models trained on NIHxPDC. Best results in **red**.

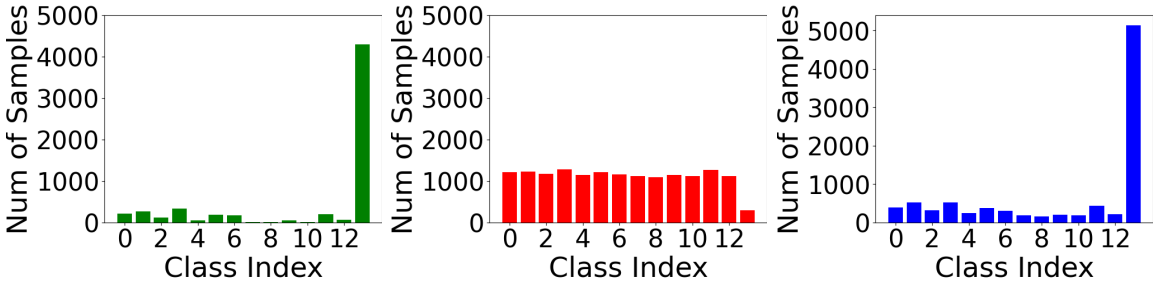
but for $p_s, p_l = 60\%$, NVUM improves 2.8% over BoMD. Hence, BoMD provides a consistently better classification result across many noise rates, but for large noise rates, our re-labelling method may be injecting too much noise into the training set. Another interesting point to note is that with 0% controlled noise, BoMD shows 89.76% AUC, which means that the AUC drops an average of 3.8% for each addition of 20% for p_s with a fixed $p_l = 20\%$.



(a) Recall (b) Precision

Figure 6.3: Noisy-label sample detection performance on NIHxPDC. We compare our proposed rank-based detection approach with DivideMix's small loss method [106] for a) recall, and b) precision. The horizontal axes show the values for $[p_s, p_l]$.

We now study the effectiveness of the detection and re-labelling of noisy samples by BoMD. In Fig. 6.3, we compare the precision and recall of our detection of noisy-label samples compared with the traditional small-loss approach used by DivideMix [106]. Notice that our noisy-label sample detection consistently outperforms DivideMix's small-loss method on both measures. An interesting note is that while recall worsens, precision improves with increasing noise rates. This happens because of the natural imbalance found in the distribution of classes in CXR datasets, where the



(a) Clean (b) Noise (c) Re-labelled

Figure 6.4: Visualisation of the changes in the histogram of label distributions after applying our re-labelling.**Left**: label distribution for the clean set OpenI. **Middle**: label distribution after injecting symmetric noise $p_s = 0.4, p_l = 0.2$. **Right**: label distribution after the re-labelling by BoMD.

class "No Findings" is dominant. A larger synthetic noise rate implies that this class is more affected than the others, but at the same time easier to detect given that the NIH dataset has relatively smaller noise rates.

We also visualise in Fig. 6.4 the label distribution change before and after applying our re-labelling. Note that our method successfully corrects the noisy label distribution to be closer to the original clean label distribution. The mean AUC over the labels in the re-labelled dataset before and after our re-labelling process is presented in Fig. 6.5, where results show that our re-labelling process significantly improves the label cleanliness of the training set for all benchmark noise rates. Recall that in multi-class problems, such re-labelling is facilitated by the fact that each sample can only have a single label. However, such constraint is dropped for multi-label problems, making the re-labelling more complicated because the feature space will be populated with multiple clusters containing different combinations of multi-labels. In the Appendix C, we evaluate the amount of consistency the KNN neighboring samples need to have for a clean re-labelling. This evaluation is based on the label-wise precision and recall results of our graph-based re-labelling method as a function of a threshold on the minimum number of nearest neighbors containing the same label. Results suggest that precision and recall increase until this threshold is between 4 and 6 nearest neighbours, and plateaus afterwards. Hence, when 4 to 6 neighbours (depending on the noisy rate) share a particular label, it is probable that the noisy training sample has this clean label.

6.4.4 Ablation Study

Language models. Our ablation study starts with an investigation of the language models for our BoMD, where we consider three types of models: 1) a randomly ini-

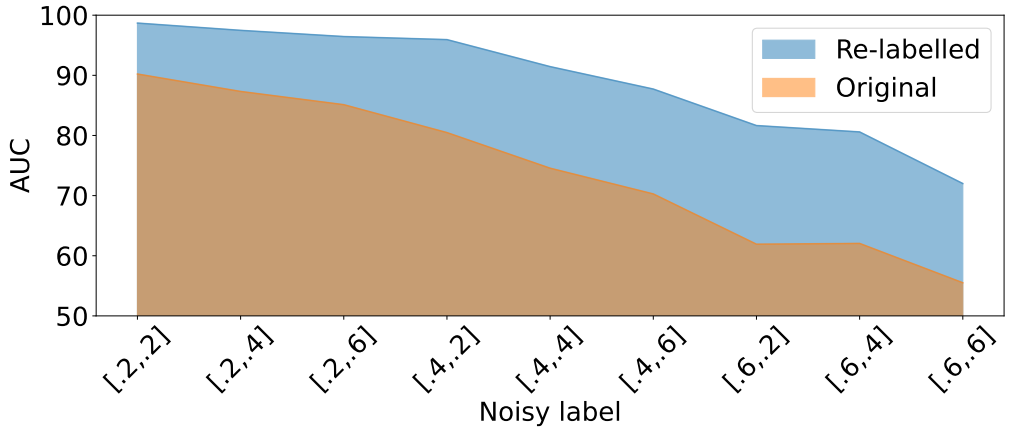


Figure 6.5: Mean AUC over labels before (orange) and after (blue) our re-labelling w.r.t PadChest's clean labels. The horizontal axes show values for $[p_s, p_l]$.

tialised model (without relying on language models); 2) a computer vision language model (e.g., Common Crawl data [4]); and 3) a medical language model (e.g., BioBERT[5], ClinicalBERT[6], BlueBERT[7]). In Table 6.5, the BERTs box shows that BERT models enable better performance than other models, with gains from 2.50% to 1.50% on OpenI and PadChest. We argue that this is because the word embeddings from BERT models contain relevant clinical semantic meaning (e.g., Atelectasis and Pneumonia are both correlated to lung opacity, but uncorrelated with Enlarged Cardiomediastinal [75]) that facilitates the multi-label descriptor learning of our method. Among the models trained on BERT models, we observe small variations, which can be related to: 1) the size of the training set, and 2) the relatedness of the medical dataset to our CXR classification problem (BlueBERT is arguably more related than BioBERT).

Evaluation of MID. In Table 6.5 the Stage-one training box shows a study of the effectiveness of MID for graph construction and downstream classification task, by comparing it against the use of the descriptors from NVUM [114]. Given that NVUM produces one descriptor per image, we set MID's number of descriptors per image at $M = 1$ for fairness. Results from the table show that MID descriptors allow an improvement of between 2% and 3% compared to NVUM's descriptors. In addition, $M = 1$ represents a graph with one (instead of multiple) descriptor per image, where we can observe a 1% to 2% drop, which indicates that our aggregating sub-graph is a more suitable strategy for multi-label images.

Smoothly re-labelling. We show an ablation study of the mixup terms in Eq. (6.6) in terms of the testing AUC results in Table 6.6. First, we mix up $\mathbf{y}$ and the uniform

[4] https://commoncrawl.org

[5] Biomedical language model, pretrained on PubMed [104].

[6] MIMIC corpus (FT on BioBERT) [74].

[7] Pretrained on PubMed abstract + MIMIC-III (clinical notes) [149].

Ablation Study	M	Language Models	Open-i	PadChest
BERTs	3	Random Init.	87.02	84.99
	3	Glove [150]	87.62	85.08
	3	ClinicalBERT [74]	88.27	85.72
	3	BioBERT [104]	89.11	86.27
	3	BlueBERT [149]	**89.52**	**86.50**
Stage-one training	1	Self-supervised [238]	84.50	83.21
	1	NVUM [114]	86.69	84.66
	1	MID	88.34	86.02
	3	MID	**89.52**	**86.50**

Table 6.5: Ablation study that compares the mean testing AUC results of our BoMD with the use of different language models (BERTs) and descriptor training (Stage-one training), M shows the number of descriptors per image.

distribution $\mathbf{1}$ (i.e., label smoothing) with a fixed mixup coefficient of $\lambda = 0.6$ (first row of Table 6.6), then we introduce $\bar{\mathbf{y}}$ with another fixed mixup coefficient of $\gamma = 0.25$ (second row of Table 6.6) and observe improvements of over 4%. Next, we remove $\mathbf{1}$ and add the binary mask $\mathbf{m}$ (third row of Table 6.6) to filter out the confident negative labels, which increases the performance by around 1%. The integration of all re-labelling components further increases performance from 0.23% to 0.41%. These results suggest that the mask $\mathbf{m}$ combined with the KNN average label $\bar{\mathbf{y}}$ mitigate the over-smoothing promoted by the uniform distribution $\mathbf{1}$.

Pre-training with vision-language model [238]. CoOp [238] is an effective visual-textual pre-training that can be considered for improving any of the general methods in Tab. 6.1 and 6.2. Hence, we pre-trained with CoOp the method in [68] and noticed a 1.04% and 0.69% performance drop in Tab. 6.1 and 6.2. This drop can be explained by the fact that CoOP requires backpropagation for the context tokens of the labels during training, where noisy labels may have caused the learned token to be inaccurate.

Additional results. In the Appendix C, we include a visualisation of different label smoothing techniques using a t-SNE [185] graph for a toy problem, additional evaluation for descriptors extracted by the MID module, and detailed sensitivity testing of hyper-parameters.

$\mathbf{1}$	$\bar{\mathbf{y}}$	$\mathbf{m}$	OpenI	PadChest
✓			83.72	80.93
✓	✓		87.92	85.48
	✓	✓	89.11	86.27
✓	✓	✓	**89.52**	**86.50**

Table 6.6: Ablation study of the testing AUC results of the components of our re-labelling in (6.6). $\bar{\mathbf{y}}$ indicates the KNN propagated label, $\mathbf{1}$ is the uniform distribution, and $\mathbf{m}$ is the binary mask.

6.5 Discussion and Conclusion

In this work, we proposed BoMD, a new method to learn from noisy multi-label CXR datasets. BoMD explores the clinical semantic information, represented by word embeddings from BlueBERT [149], to optimise the multi-label image descriptors which are used to find noisy multi-label training samples. We then use the learned image descriptors to build a graph for smoothly re-labelling the training data. BoMD outperforms current SOTA methods on three real-world CXR benchmarks that consist of training on two large-scale noisy multi-label CXR datasets and testing on three clean multi-label CXR datasets. We additionally evaluate BoMD on our proposed systematic benchmark to further show the effectiveness and robustness of our method.

Limitations and future work. We identify three limitations of BoMD. The first issue is the longer training time (+8h compared with NVUM [114]) since it requires multiple training stages. We plan to tackle this problem by better integrating the training stages. The second issue is that BoMD decreases its performance under extremely noisy label setup (i.e., [0.6, 0.6] in Table 6.4), which is due to mistakes in the smooth re-labelling. However, such a high noise rate may not be applicable in real-world scenarios since the usual F1 score for text mining performance is between 80% to 94% [38, 143, 193], which suggests noise rates much smaller than 60%. Another drawback of BoMD is that it does not address imbalanced learning, which is an important point when training with CXR datasets. The future work will study how to combine first stage method with imbalanced learning approach such as logit adjustment [135]. BoMD demonstrates the possibility of leveraging semantic information and sample graphs to estimate the label distribution for training a better CXR classifier. However, noisy-cleaning methods are still dominant in the LNL under multi-class scenarios. One potential future direction is to study a unified framework for addressing both multi-class and multi-label tasks with minimum adaptation.

6.6 Acknowledgement

This work was supported by funding from the Australian Research Council through grant FT190100525.

Statement of Authorship

Title of Paper	Asymmetric Co-teaching with Multi-view Consensus for Noisy Label Learning
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted work written in manuscript style
Publication Details	

Principal Author

Name of Principal Author (Candidate)	Fengbei Liu
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper and revision
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: 09/14/2023

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 in 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 and revision
Signature	Date: 09/15/2023

Name of Co-Author	Chong Wang		
Contribution to the Paper	Discussion and commented on the revision		
Signature		Date	

Name of Co-Author	Yu Tian		
Contribution to the Paper	Discussion and commented on the revision		
Signature		Date	

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	Discussion and wrote the revision		
Signature		Date	23/09/2023

Chapter 7

AsyCo: Asymmetric Co-teaching with Multi-view Consensus for Noisy Label Learning

Abstract

Learning with noisy-labels has become an important research topic in computer vision where state-of-the-art (SOTA) methods explore: 1) prediction disagreement with co-teaching strategy that updates two models when they disagree on the prediction of training samples; and 2) sample selection to divide the training set into clean and noisy sets based on small training loss. However, the quick convergence of co-teaching models to select the same clean subsets combined with relatively fast overfitting of noisy labels may induce the wrong selection of noisy label samples as clean, leading to a confirmation bias that damages accuracy. In this paper, we introduce our noisy-label learning approach, called Asymmetric Co-teaching (AsyCo), which introduces a new prediction disagreement strategy that consistently produce divergent predictions for noisy samples by co-teaching models, and a new sample selection approach that does not require the small-loss assumption to enable better robustness to confirmation bias than previous methods. More specifically, the new prediction disagreement is achieved with the use of different training strategies, where one model is trained with multi-class learning and the other with multi-label learning. Also, the new sample selection is based on multi-view consensus, which uses the label views from training labels and model predictions to divide the training set into clean and noisy for training the multi-class model and to re-label the training samples with multiple top-ranked labels for training the multi-label model. Extensive experiments on synthetic and real-world noisy-label datasets show that AsyCo improves over current SOTA methods.

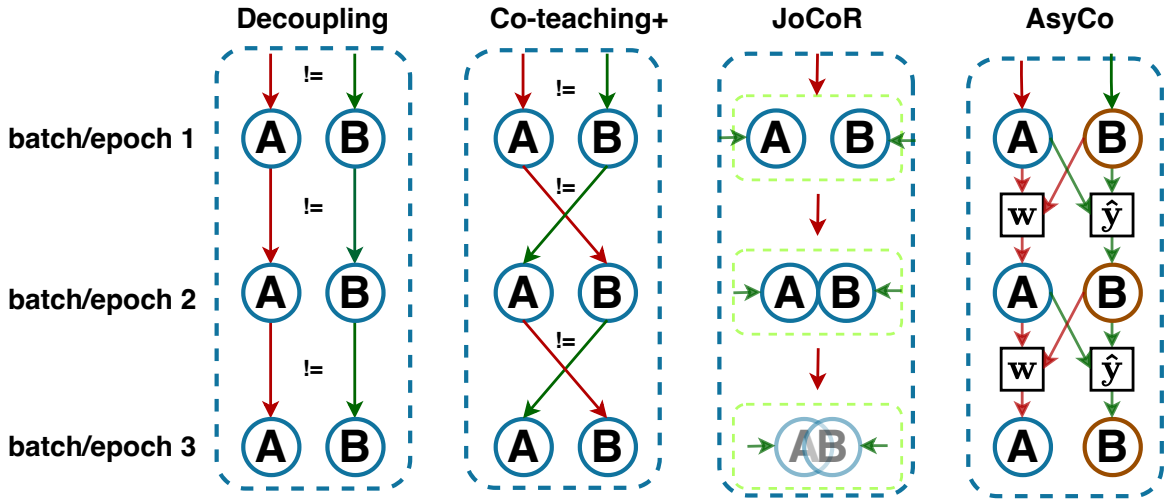


Figure 7.1: Diagrams of the following noisy-label learning methods: Decoupling [134], Co-teaching+ [224], JoCoR [198], and our AsyCo. AsyCo co-teaches the multi-class model A and the multi-label model B with different training strategies (denoted by the different colours of A&B). The training samples for A and B, represented by the green and red arrows, are formed by our proposed multi-view consensus that uses label views from the training set and model predictions to estimate the variables $\mathbf{w}$ and $\hat{\mathbf{y}}$, which selects clean/noisy samples for training A and iteratively re-labels samples for training B, respectively.

7.1 Introduction

Deep neural network (DNN) has achieved remarkable success in many fields, including computer vision [67, 97], natural language processing (NLP) [41, 223] and medical image analysis [113, 192]. However, the methods from those fields often require massive amount of high-quality annotated data for supervised training [40], which is challenging and expensive to acquire. To alleviate such problem, some datasets have been annotated via crowdsourcing [211], from search engines [171], or with NLP from radiology reports [192]. Although these cheaper annotation processes enable the construction of large-scale datasets, they inevitably introduce noisy labels for model training, resulting in DNN model performance degradation. Therefore, novel learning algorithms are required to robustly train DNN models when training sets containing noisy labels.

Previous methods tackle noisy-label learning from different perspectives. For example, some approaches focus on *prediction disagreement* [134, 198, 224], which rely on jointly training two models to update their parameters when they disagree on the predictions of the same training samples. These two models generally use the same training strategy, so even though they are trained using samples with divergent predic-

tions, both models will quickly converge to select similar clean samples during training, which neutralises the effectiveness of prediction disagreement. Other noisy-label learning methods are based on *sample selection* [1, 60, 106] to find clean and noisy-label samples that are treated differently in the training process. Sample-selection approaches usually assume that samples with small training losses are associated with clean labels, which is an assumption verified at early training stages [119, 226]. However, such assumption is unwarranted in later training stages because DNN models can overfit any type of noisy label after a certain number of epochs, essentially reducing the training loss for all training samples. State-of-the-art (SOTA) noisy-label learning approaches [106] have been designed to depend on both prediction disagreement and sample selection methods to achieve better performance than either method alone. Nevertheless, these SOTA methods are still affected by the fast convergence of both models and label noise overfitting, which raises the following questions: 1) **Are there more effective ways to keep the training of both models divergent for noisy samples, so they do not easily converge during the training procedure? 2) Is there a sample selection approach that can be better integrated with prediction disagreements than the small loss strategy?**

Motivated by traditional multi-view learning [13, 165] and multi-label learning, we propose a new noisy-label learning method that aims to answer the two questions above. Our method, named **Asymmetric Co-teaching (AsyCo)** and depicted in Fig. 7.1, is based on two models trained with different learning strategies to consistently produce divergent predictions for noisy samples. One model, the **classification net**, is trained with conventional multi-class learning by minimising a cross entropy loss and provide single-class prediction, and the other, the **reference net**, is trained with a binary cross entropy loss to enable multi-label learning that is used to estimate the top-ranked labels that represent the potentially clean candidate labels for each training sample. The original training labels and the predictions by the training and reference nets enable the formation of three label views for each training sample, allowing us to formulate the **multi-view consensus** that is tightly integrated with the prediction disagreement to select clean and noisy samples for training the multi-class model and to iteratively re-label samples with multiple top-ranked labels for training the multi-label model. In summary, our main contributions are:

- The new noisy-label co-teaching method **AsyCo** designed to consistently produce divergent predictions for noisy samples by applying different training strategies for two models trained with noisy-label samples.

- The novel **multi-view consensus** that uses the disagreements between training labels and model predictions to select clean and noisy samples for training the multi-class model and to iteratively re-label samples with multiple top-ranked labels for training the multi-label model.

We conduct extensive experiments on both synthetic and real-world noisy datasets that show that AsyCo provides substantial improvements over previous SOTA methods.

7.2 Related Work

Prediction disagreement approaches seek to maximise model performance by exploring the prediction disagreements between models trained from the same training set. In general, these methods [82, 134, 198, 224] train two models using samples that have different predictions from both models to mitigate the problem of confirmation bias (i.e., a mistake being reinforced by further training from the same mistake) that particularly affects single-model training. Furthermore, the cross teaching of two models can help escape local minima. Most of the prediction-disagreement methods also rely on sample-selection techniques, as we explain below, but in general, they use the same training strategy to train two models, which limits the ability of these approaches to delay the convergence between the models.

Sample selection approaches aim to automatically classify training samples into clean or noisy and treat them differently during the training process. Previous papers [119, 226] have shown that when training with noisy label, DNN fits the samples with clean labels first and gradually overfits the samples with noisy labels later. Such training loss characterisation allowed researchers to assume that samples with clean labels have small losses, particularly at early training stages – this is known as the *small-loss assumption*. For examples, M-correction [1] automatically selects clean samples by modelling the training loss distribution with a Beta Mixture model (BMM). Sample selection has been combined with prediction disagreement in several works, such as Co-teaching [60] and Co-teaching+ [224] that train two networks simultaneously, where in each mini-batch, it selects small-loss samples to be used in the training of the other model. JoCoR [198] improves upon Co-teaching+ by using a contrastive loss to jointly train both models. DivideMix [106] has advanced the area with a similar combination of sample selection and prediction disagreement using semi-supervised learning, co-teaching and small-loss detection with a Gaussian Mixture Model (GMM). InstanceGM [48] combines graphical model with DivideMix to achieve promising results. These methods show that sample selection based on the small-loss assumption is one of the core components for achieving SOTA performance. However, the small loss signal used to select samples is poorly integrated with prediction disagreement since both models will quickly converge to produce similar loss values for all training samples, resulting in little disagreement between models and increase confirmation bias risk.

Multi-view learning (MVL) studies the integration of knowledge from different views of the data to capture consensus and complementary information across different views. Traditional MVL methods [13, 165] aimed to encourage the convergence of

patterns from different views. For example, Co-training [13] uses two views of web-pages (i.e., text and hyperlinks on web-pages) to allow the use of inexpensive unlabelled data to augment a small labelled data. Recent methods [61] weight the contribution of each view based on the estimated uncertainty. In our paper, we explore this multi-view learning strategy to select clean and noisy samples and to iteratively re-label training samples, where the views are represented by the training labels, and the predictions by the two models with different learning strategies.

7.2.1 Problem Definition

We denote the noisy training set as $\mathcal{D} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^{H \times W \times C}$ is the input image of size $H \times W$ with C colour channels, and $\tilde{\mathbf{y}}_i \in \mathcal{Y} \subset \{0,1\}^{|\mathcal{Y}|}$ is the one-hot (or multi-class) label representation. Our main goal is to learn the **classification net** $n_\theta : \mathcal{X} \to \mathcal{L}$, parameterised by $\theta \in \Theta$, that outputs the logits $\mathbf{l} \in \mathcal{L} \subset \mathbb{R}^{|\mathcal{Y}|}$ for an image $\mathbf{x} \in \mathcal{X}$. Our prediction-disagreement strategy requires the definition of the **reference net** denoted by $r_\phi : \mathcal{X} \to \mathcal{L}$, parameterised by $\phi \in \Phi$, to be jointly trained with $n_\theta(.)$.

AsyCo[1] is based on alternating the training of the multi-class model $n_\theta(.)$ and the multi-label model $r_\phi(.)$, which allows the formation of three label views for the training samples $\mathbf{x}_i$ in $\mathcal{D}$, namely: 1) the original training label $\tilde{\mathbf{y}}_i$, 2) the classification net multi-class prediction $\tilde{\mathbf{y}}_i^{(n)}$, and 3) the reference net multi-label prediction $\tilde{\mathbf{y}}_i^{(r)}$. Using these views, we introduce new methods to estimate the sample-selection variable $\mathbf{w} \in \mathbb{R}^{|\mathcal{D}|}$ that classifies training samples into clean or noisy, and the re-labelling variable $\hat{\mathbf{y}} \in [0,1]^{|\mathcal{D}| \times |\mathcal{Y}|}$ that holds multiple top-ranked labels for training samples, where $\mathbf{w}$ is used for training the multi-class model $n_\theta(.)$, and $\hat{\mathbf{y}}$ for training the multi-label model $r_\phi(.)$. Fig. 7.2 depicts AsyCo, in comparison with prediction disagreement methods based on co-teaching and small-loss sample selection.

7.2.2 Asymmetric training

Our AsyCo optimisation starts with a warmup stage of supervised learning to train each network with different losses:

$$\begin{aligned} \theta &= \arg\min_\theta \sum_{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \mathcal{D}} \ell_{\text{CE}}(\tilde{\mathbf{y}}_i, \sigma_{sm}(n_\theta(\mathbf{x}_i))), \\ \phi &= \arg\min_\phi \sum_{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \mathcal{D}} \ell_{\text{BCE}}(\tilde{\mathbf{y}}_i, \sigma_{sg}(r_\phi(\mathbf{x}_i))), \end{aligned} \tag{7.1}$$

where $\sigma_{sm}(.)$ and $\sigma_{sg}(.)$ are the softmax and sigmoid activation functions, respectively, $\ell_{\text{CE}}(.)$ represents the CE loss for multi-class learning, and $\ell_{\text{BCE}}(.)$ denotes the BCE loss

[1] Algorithm in Appendix D.

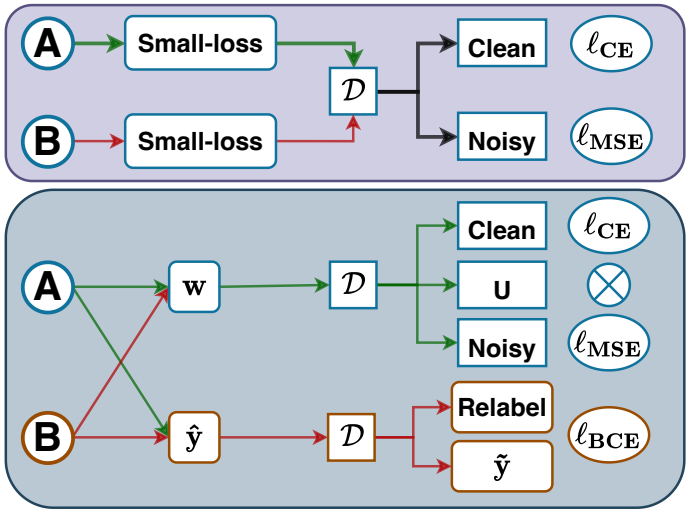


Figure 7.2: Comparison between traditional small-loss sample selection (top) and our AsyCo, consisting of prediction disagreement between the multi-class model A and multi-label model B (bottom). Traditional methods utilise the small-loss assumption for classifying samples as clean or noisy, while our multi-view sample selection uses prediction disagreements to update the sample-selection variable $\mathbf{w}$ for classifying samples as clean, noisy or unmatched (U) to train the classification net A. Our multi-view re-labelling enforces model disagreement to update the re-labelling variable $\hat{\mathbf{y}}$ for selecting ambiguous samples to be re-labelled for reference net B.

for multi-label learning. The two models from (7.1) will provide predictions as follows:

$$\begin{aligned} \tilde{\mathbf{y}}_i^{(n)} &= \text{OneHot}(n_{\theta^\dagger}(\mathbf{x}_i)), \\ \tilde{\mathbf{y}}_i^{(r)} &= \text{TopK}(r_{\phi^\dagger}(\mathbf{x}_i)), \end{aligned} \tag{7.2}$$

where $\tilde{\mathbf{y}}_i^{(n)} \in \mathcal{Y}$ is the one-hot single-label prediction by $n_{\theta^\dagger}(\mathbf{x}_i)$ (i.e., the largest value from $n_{\theta^\dagger}(.)$ will set $\tilde{\mathbf{y}}_i^{(n)}$ to 1 and the rest are set to 0), and $\tilde{\mathbf{y}}_i^{(r)} \in \{0,1\}^{|\mathcal{Y}|}$ is the top-K multi-label prediction of $r_{\phi^\dagger}(\mathbf{x}_i)$ (i.e., the largest K values from $r_{\phi^\dagger}(.)$ will set $\tilde{\mathbf{y}}_i^{(r)}$ to 1 and the rest are set to 0).

Note that although both $\ell_{\text{CE}}(.)$ and $\ell_{\text{BCE}}(.)$ are optimised with $\tilde{\mathbf{y}}$, the supervisory signal of the $\ell_{\text{BCE}}(.)$ is different from $\ell_{\text{CE}}(.)$ since it tries to optimise each label direction independently. This is helpful when $\tilde{\mathbf{y}}$ is noisy as the network will still receive correct supervisory signal from most of the negative labels, where the clean label may not be top prediction but in the top-K prediction set. Nevertheless, this also removes the single-label constraint for multi-class learning, which will weaken the training convergence. Therefore, the multi-label network is not used for inference, but we aim to extract useful information from its top-ranked labels to help with the training of $n_\theta(.)$ using the multi-view consensus[2]. As explained below, our training uses the label views produced by the predictions from $n_\theta(.)$ and $r_\phi(.)$ and the training labels, to select samples for training $n_\theta(.)$ and to re-label samples for training $r_\phi(.)$.

7.2.3 Multi-view Consensus

One of the objectives of utilizing prediction disagreement between models is to improve sample selection accuracy for co-teaching. We propose a new sample selection based on multi-view consensus, where each sample $\mathbf{x}_i$ has three label views, i.e., the training label $\tilde{\mathbf{y}}_i$, the single-label one-hot prediction $\tilde{\mathbf{y}}_i^{(n)}$, and the multi-label top-K prediction $\tilde{\mathbf{y}}_i^{(r)}$. These multiple views allow us to build training subsets given prediction disagreements, as shown in Tab. 7.1 The **training of the classification net** $n_\theta(.)$ has the goals of producing the testing model and classifying training samples in terms of the disagreement with $r_\phi(.)$ to divide the training set into clean and noisy label samples. Unlike previous methods that rely on the small-loss assumption to classify training samples into clean or noisy [1, 60, 106], we utilize the subsets created by prediction disagreements from the multiple label views shown in Tab. 7.1. For training $n_\theta(.)$, we seek label agreements between the pair of views *beyond* its own prediction. More specifically, for $n_\theta(.)$, training samples are classified as clean when $\tilde{\mathbf{y}}^\top \tilde{\mathbf{y}}^{(r)} = 1$, which indicates that the training label matches one of the top-ranked predictions by $r_\phi(.)$. Such agreement from label views $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{y}}^{(r)}$, established by the other pair of views, can help reduce the confirmation bias from $n_\theta(.)$'s prediction, and at the same time

[2]Training strategy visualization in Appendix D.

Table 7.1: The three possible label views are the training label $\tilde{\mathbf{y}}_i$, the single-label one-hot prediction $\tilde{\mathbf{y}}_i^{(n)}$, and the multi-label top-K prediction $\tilde{\mathbf{y}}_i^{(r)}$. The combination of these multiple views form the subsets listed in this table, where "1" means the two views agree and "0" means the two views disagree.

Subsets	$\tilde{\mathbf{y}}^\top \tilde{\mathbf{y}}^{(n)}$	$\tilde{\mathbf{y}}^{(n)\top} \tilde{\mathbf{y}}^{(r)}$	$\tilde{\mathbf{y}}^\top \tilde{\mathbf{y}}^{(r)}$
Core (C)	1	1	1
Side-Core (SC)	0	1	1
$(n_\theta, \tilde{\mathbf{y}})$ (NY)	1	0	0
(n_θ, r_ϕ) (NR)	0	1	0
$(r_\phi, \tilde{\mathbf{y}})$ (RY)	0	0	1
Unmatched (U)	0	0	0

extract useful labelling information from $r_\phi(.)$ to improve $n_\theta(.)$'s performance. On the other hand, the training samples, where $\tilde{\mathbf{y}}^\top \tilde{\mathbf{y}}^{(r)} = 0$ but $\tilde{\mathbf{y}}_i^\top \tilde{\mathbf{y}}_i^{(n)} + \tilde{\mathbf{y}}_i^{(n)\top} \tilde{\mathbf{y}}_i^{(r)} = 1$, are classified as problematic because the label $\tilde{\mathbf{y}}$ is not in the top-ranked predictions from $\tilde{\mathbf{y}}^{(r)}$. We can further sub-divide this category of samples into noisy and unmatched, with the former containing samples where $\tilde{\mathbf{y}}^{(n)}$ appears to be clean (for which we use a robust loss function, e.g., MSE) and the latter has samples that none of the views match with each other (these samples are discarded from training).

Therefore, based on the criterion described above and the subsets from Tab. 7.1, the classification net $n_\theta(.)$ is trained with $\{\mathrm{C}, \mathrm{SC}, \mathrm{RY}\}$ as clean, $\{\mathrm{NY}, \mathrm{NR}\}$ as noisy and discard $\{U\}$, defined by the following sample-selection variable:

$$\mathbf{w}_i = \begin{cases} +1, & \text{if } \tilde{\mathbf{y}}_i^\top \tilde{\mathbf{y}}_i^{(r)} = 1, \\ 0, & \text{if } \tilde{\mathbf{y}}_i^\top \tilde{\mathbf{y}}_i^{(r)} = 0 \text{ and } \tilde{\mathbf{y}}_i^\top \tilde{\mathbf{y}}_i^{(n)} + \tilde{\mathbf{y}}_i^{(n)\top} \tilde{\mathbf{y}}_i^{(r)} = 1, \\ -1, & \text{if } \tilde{\mathbf{y}}_i^\top \tilde{\mathbf{y}}_i^{(r)} = 0 \text{ and } \tilde{\mathbf{y}}_i^\top \tilde{\mathbf{y}}_i^{(n)} + \tilde{\mathbf{y}}_i^{(n)\top} \tilde{\mathbf{y}}_i^{(r)} = 0, \end{cases} \tag{7.3}$$

where $\mathbf{w}_i \in \{+1, 0, -1\}$ denotes a clean, noisy, and unmatched training sample, respectively.

The training of $n_\theta(.)$ is performed by

$$\begin{aligned} \theta^* = \arg\min_\theta & \sum_{\substack{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \mathcal{D} \\ \mathbf{w}_i = +1}} \ell_{CE}(\tilde{\mathbf{y}}_i, \sigma_{sm}(n_\theta(\mathbf{x}_i))) \\ & + \lambda \sum_{\substack{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \mathcal{D} \\ \mathbf{w}_i = 0}} \ell_{MSE}(\upsilon(\sigma_{sm}(n_\theta(\mathbf{x}_i)), T), \sigma_{sm}(n_\theta(\mathbf{x}_i))), \end{aligned} \tag{7.4}$$

where $\upsilon(., T)$ is a temperature sharpening function parameterised by T, and λ is the weight to control the strength of the unsupervised learning with its own prediction, and

$\ell_{MSE}(.)$ denotes the mean square error loss function. The **training of the reference net** $r_\phi(.)$ focuses on enforcing prediction disagreement for both models. This optimisation goal focuses on designing new supervisory training signals that temporarily re-label the samples where predictions by $n_\theta(.)$ and $r_\phi(.)$ match (i.e., $\tilde{\mathbf{y}}^{(n)\top}\tilde{\mathbf{y}}^{(r)} = 1$) but where $n_\theta(.)$ does not match the training label $\tilde{\mathbf{y}}$ (i.e., $\tilde{\mathbf{y}}^\top\tilde{\mathbf{y}}^{(n)} = 0$). The training samples that meet this condition can be regarded as hard to fit by $n_\theta(.)$, with the top-ranked predictions by $\tilde{\mathbf{y}}^{(r)}$ being likely to contain the hidden clean label. The conditions above indicate that we select samples from $\{\text{SC}, \text{NR}\}$ from Table 7.1 for re-labelling. For samples in SC, since $n_\theta(.)$ is trained with supervised learning with $\tilde{\mathbf{y}}$ in (7.4), we re-label the sample to $\tilde{\mathbf{y}}^{(n)}$. For samples in NR, $n_\theta(.)$ is trained with unsupervised learning in (7.4), so we re-label the sample to $\tilde{\mathbf{y}} + \tilde{\mathbf{y}}^{(n)}$, forming a multi-label target. We define the re-labelling variable $\hat{\mathbf{y}}$ to represent the new supervisory training signal, as follows:

$$\hat{\mathbf{y}}_i = \begin{cases} \tilde{\mathbf{y}}_i^{(n)}, & \text{if } (\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \text{SC}, \\ \tilde{\mathbf{y}}_i + \tilde{\mathbf{y}}_i^{(n)}, & \text{if } (\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \text{NR}, \\ \tilde{\mathbf{y}}_i, & \text{otherwise.} \end{cases} \tag{7.5}$$

The training of $r_\phi(.)$ is achieved with:

$$\phi^* = \arg\min_\phi \sum_{i=1}^{|\mathcal{D}|} \ell_{BCE}(\hat{\mathbf{y}}_i, \sigma_{sg}(r_\phi(\mathbf{x}_i))). \tag{7.6}$$

Note that this re-labelling is iteratively done at every epoch. The testing procedure depends exclusively on the classification net $n_\theta(.)$.

7.3 Experiments

We show the results of extensive experiments on instance-dependent synthetic noise benchmarks with datasets CIFAR10 and CIFAR100 [96] with various noise rates and on three real-world datasets, namely: Animal-10N [171], Red Mini-ImageNet [79] and Clothing1M [211].

7.3.1 Datasets

CIFAR10/100. For CIFAR10 and CIFAR100 [96], the training set contains 50K images and testing set contains 10K images of size $32 \times 32 \times 3$. CIFAR10 has 10 classes and CIFAR100 has 100 classes. We follow previous work [209] for generating instance-dependent noise (IDN) with rates in $\{0.2, 0.3, 0.4, 0.5\}$. **Red Mini-ImageNet** is proposed by [79] based on Mini-ImageNet [40]. The images and their corresponding labels are annotated by Google Cloud Data Labelling Service. This dataset is proposed to study real-world web-based noisy label. Red Mini-ImageNet has 100 classes with

each class containing 600 images from ImageNet. The images are resized to 32×32 from the original 84×84 pixels to allow a fair comparison with other baselines [79, 216]. We test our method on noise rates in {20%, 40%, 60%, 80%}. **Animal 10N** is a real-world dataset proposed in [171], which contains 10 animal species with similar appearances (wolf and coyote, hamster and guinea pig, etc.). The training set size is 50K and testing size is 10K, where we follow the same setup as [171]. **Clothing 1M** is a real-world dataset with 100K images and 14 classes. The labels are generated from surrounding text with an estimated noise ratio of 38.5%. The dataset also contains clean training, clean validation and clean test sets with 50K, 14K and 10K images. We do not use clean training and clean validation, only the clean testing is used for measuring model performance.

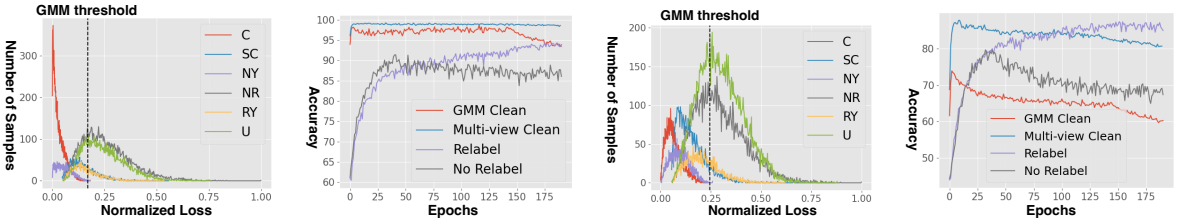
7.3.2 Implementation

For the implementation, we use the baseline models that most papers use for each dataset. For CIFAR10/10 and Red Mini-ImageNet we use Preact-ResNet18 [67] and train it for 200 epochs with SGD with momentum=0.9, weight decay=5e-4 and batch size=128. The initial learning rate is 0.02 and reduced by a factor of 10 after 150 epochs. The warmup period for all three datasets is 10 epochs. We set $\lambda = 25$ in (7.4) for CIFAR10 and Red Mini-ImageNet, and $\lambda = 100$ for CIFAR100. In (7.2), we set $K = 1$ for CIFAR10 and $K = 3$ for CIFAR100 and Red Mini-ImageNet. These values are fixed for all noise rates. In Appendix D studies λ and K on CIFAR100, and results show strong robustness to a range of values for these variables For data augmentations, we use random cropping and random horizontal flipping for all three datasets.

For Animal 10N, we follow the setup used by previous methods with a VGG-19BN [164] architecture, trained for 100 epochs with SGD with momentum=0.9, weight decay=5e-4 and batch size=128. The initial learning rate is 0.02, and reduced by a factor of 10 after 50 epochs. The warmup period is 10 epochs. We set $\lambda = 25$ and $K = 2$. For data augmentations, we use random cropping and random horizontal flipping.

For Clothing1M, we follow the common setup that uses the ImageNet [40] pre-trained ResNet50 [67] and train it for 40 epochs with SGD with momentum=0.9, weight decay=1e-3 and batch size=32. The warmup period is 1 epoch. The initial learning rate is set to 0.002 and reduced by a factor of 10 after 20 epochs. Following DivideMix [106], we also sample 1000 mini-batches from the training set to ensure the training set is pseudo balanced. We set $K = 4$. For data augmentation, we first resize the image to 256×256 pixels, then random crop to 224×224 and random horizontal flipping.

For the semi-supervised training of $n_\theta(.)$, we use MixMatch [12] from DivideMix [106]. We also extend our method to train two $n_\theta(.)$ models and use ensemble prediction at inference time, similarly to DivideMix [106]. We denoted this variant as $2 \times n_\theta$. Our code is implemented in Pytorch [147] and all experiments are performed on an RTX



(a) CIFAR100 0.2 loss (b) CIFAR100 0.2 Accuracy (c) CIFAR100 0.5 loss (d) CIFAR100 0.5 Accuracy

Figure 7.3: (a) and (c) are sample loss histograms for the subsets in Tab. 7.1 for CIFAR100 with 0.2 and 0.5 instance-dependent noise after warmup. Vertical dot line is GMM threshold. (b) and (d) show the accuracy of the clean set selected by GMM and our multi-view strategy. (b) and (d) also show the accuracy of whether the hidden clean label is within $r_\phi(.)$'s top-ranked prediction for multi-view re-labelling compared with not using any re-labelling.

Model	Ablation	CIFAR10				CIFAR100			
		0.2	0.3	0.4	0.5	0.2	0.3	0.4	0.5
n_θ	$\mathbf{w}_i = 0$ if $(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in$ RY	93.28	93.85	92.54	82.60	73.58	71.51	65.51	56.65
	$\mathbf{w}_i = 0$ if $(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in$ U	95.71	94.88	94.34	91.60	75.10	72.64	67.42	57.55
	$\mathbf{w}_i = +1$ if $(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in$ U	95.20	95.14	94.72	90.27	75.34	73.21	66.09	55.95
	Small-loss subsets	92.37	91.80	90.93	78.53	70.10	69.52	64.69	56.35
r_ϕ	CE	95.22	94.83	83.48	64.96	73.33	69.29	63.82	54.83
	Frozen after warmup	91.19	88.97	84.72	67.57	68.73	65.36	58.88	48.13
	$\hat{\mathbf{y}}_\mathbf{i} = \tilde{\mathbf{y}}_\mathbf{i}$	95.42	94.69	90.53	84.95	74.43	71.75	62.25	53.69
	$\hat{\mathbf{y}}_\mathbf{i} = \tilde{\mathbf{y}}_\mathbf{i}^{(\mathbf{n})}$	94.29	94.23	94.13	93.67	74.55	73.71	68.21	57.84
AsyCo original result:		**96.00**	**95.82**	**95.01**	**94.13**	**76.02**	**74.02**	**68.96**	**60.35**

Table 7.2: Empircal analysis for the classification net n_θ and reference net r_ϕ.

3090[3]

7.3.3 Empirical Analysis

Before presenting comparative results with the SOTA, we show an analysis of our proposed AsyCo. To understand the training dynamics of AsyCo, we visualise the training losses of subsets from Table 7.1 that are used by our multi-view consensus approach. We also provide a comparison between the small-loss sample selection and our multi-view sample selection. Then we test alternative approaches for multi-view sample selection and re-labelling. We perform all these experiments on the IDN CIFAR10/100 [209].

Loss histograms. Fig. 7.3a and Fig. 7.3c show the loss histograms after warmup for

[3]Time of Different sample selection comparison in Appendix D.

each subset in Table 7.1. To compare with small-loss sample selection approaches, we adopt the sample-selection approach by DivideMix [106] that is based on a Gaussian Mixture Model (GMM) to divide the training set into clean and noisy subsets (the vertical black dotted line is the threshold estimated by DivideMix). These graphs show that the subsets' loss histograms are relatively consistent for different noise rates. Specifically, the subset C always has the smallest loss values among all subsets, which shows that our multi-view sample selection is able to confidently extract clean samples. We also observe that NY has small loss values in both graphs. However, using NY as clean set does not produce promising performance, as shown in Table 7.2, row 'Small-loss subsets', which represents the use of almost all samples in C and NY as clean samples (since they are on the left-hand side of the GMM threshold). This indicates that the small-loss samples in NY are likely to contain overfitted noisy-label samples, whereas our multi-view sample selection successfully avoids selecting these samples.

Accuracy of clean subset and re-labelling. In Fig. 7.3b and Fig. 7.3d, we show the accuracy of the clean set selected by the GMM-based small-loss strategy and by our multi-view consensus during the training stages. We observe that multi-view selection performs consistently better than GMM in both high and low noise rates. We also validate the accuracy of the hidden clean label produced by the top ranked predictions of $r_\phi(.)$ by comparing the re-labelling produced by Eq. (7.5) versus no re-labelling (i.e., train $r_\phi(.)$ with the original training labels.) We observe that without re-labelling, the accuracy of top-ranked predictions will drop due to over-fitting in later training epochs. However, the re-labelling accuracy consistently improves, which suggests that our multi-view re-labelling consistently improves the label accuracy over time.

Different $n_\theta(.)$ training. Table 7.2 shows a study on the selection of different subsets from Table 7.1 for the sample-selection when training the classification net $n_\theta(.)$. First, we test the importance of classifying the samples in RY as clean for training $n_\theta(.)$ by, instead, treating these samples as noisy in Eq. (7.4) (i.e., by setting $\mathbf{w}_i = 0$). This new sample selection causes a large drop in performance for all cases, which suggests that RY contains informative samples that are helpful for training $n_\theta(.)$. Second, we test whether using the unmatched samples in U can improve model training, where we include them as clean or noisy samples by setting $\mathbf{w}_i = +1$ or 0, respectively. Both studies lead to worse results compared to the original AsyCo that discards U samples (see last row). Despite this result, we also notice that in low noise rates (0.2, 0.3), treating U as clean leads to slightly better accuracy than treating U as noisy. These results suggest that the high uncertainty and lack of view agreements by the samples in U lead to poor supervisory training signal, which means that discarding these samples is currently the best option. Finally, the histograms of Fig. 7.3 indicate that NY also contains small-loss samples. Therefore, we adapt traditional small-loss assumption to train our AsyCo and use the subsets C and NY as clean and treat the other subsets as noisy. As shown in the "Small-loss subset" row of Table 7.2, the

Methods	CIFAR10				CIFAR100			
	0.2	0.3	0.4	0.5	0.2	0.3	0.4	0.5
CE	75.81	69.15	62.45	39.42	30.42	24.15	21.34	14.42
Mixup [229]	73.17	70.02	61.56	48.95	32.92	29.76	25.92	21.31
Forward [148]	74.64	69.75	60.21	46.27	36.38	33.17	26.75	19.27
T-Revision [210]	76.15	70.36	64.09	49.02	37.24	36.54	27.23	22.54
Reweight [120]	76.23	70.12	62.58	45.46	36.73	31.91	28.39	20.23
PTD-R-V [209]	76.58	72.77	59.50	56.32	65.33	64.56	59.73	56.80
Decoupling [134]	78.71	75.17	61.73	50.43	36.53	30.93	27.85	19.59
Co-teaching [60]	80.96	78.56	73.41	45.92	37.96	33.43	28.04	23.97
MentorNet [82]	81.03	77.22	71.83	47.89	38.91	34.23	31.89	24.15
CausalNL [220]	81.79	80.75	77.98	78.63	41.47	40.98	34.02	32.13
JoCoR [198]	89.30	85.54	80.87	64.11	67.87	65.73	61.64	57.75
CAL [241]	92.01	-	84.96	-	69.11	-	63.17	-
kMEIDTM [26]	92.26	90.73	85.94	73.77	69.16	66.76	63.46	59.18
DivideMix [106] $\theta^{(1)}$ test †	94.62	94.49	93.50	89.07	74.43	73.53	69.18	57.52
Ours	96.00	95.82	95.01	94.13	76.02	74.02	68.96	60.35
DivideMix [106] †	94.80	94.60	94.53	93.04	77.07	76.33	70.80	58.61
Ours $2\times n_\theta$ test	96.56	96.11	95.53	94.86	78.50	77.32	73.32	65.96

Table 7.3: Test accuracy (%) of different methods on CIFAR10/100 with instance-dependent noise [209]. Results reproduced from publicly available code are presented with †. Best single/ensemble inference results are labelled with red/green.

accuracy is substantially lower, which suggests that the small-loss samples may contain overfitted noisy samples.

Different $r_\phi(.)$ training. We analyse the training of $r_\phi(.)$ with different training losses and re-labelling strategies in Table 7.2. We first study how the multi-label training loss provided by the BCE loss helps mitigate label noise by training our reference net $r_\phi(.)$ with the CE loss $\ell_{CE}(.)$ in Eq. (7.1) and (7.6), while keeping the multi-view sample selection and re-labelling strategies unchanged. We observed that by training $r_\theta(.)$ with $\ell_{CE}(.)$ leads to a significant drop in accuracy for most cases, where for CIFAR10 with low noise rate (20% and 30%), $\ell_{CE}(.)$ maintains the accuracy of $\ell_{BCE}(.)$, but for larger noise rates, such as 40% and 50%, $\ell_{CE}(.)$ is not competitive with $\ell_{BCE}(.)$ because it reduces the prediction disagreements between $n_\theta(.)$ and $r_\phi(.)$, facilitating the overfitting to the same noisy-label samples by both models. For CIFAR100, $\ell_{CE}(.)$ leads to worse results than $\ell_{BCE}(.)$ for all cases. These results suggest that to effectively co-teach two models with prediction disagreement, the use of different training strategies is an important component. Next, we study a training, where $r_\phi(.)$ is frozen after warmup, but we still train $n_\theta(.)$. The result drops significantly which indicates that $r_\phi(.)$ needs to be trained in conjunction with $n_\theta(.)$ to achieve reasonable performance. We study different re-labelling strategies by first setting $\hat{\mathbf{y}}_i = \tilde{\mathbf{y}}$ for training

$r_\phi(.)$, which leads to comparable results for low noise rates, but worse results for high-noise rates, suggesting that that only training with $\tilde{\mathbf{y}}$ is not enough to achieve good performance. Finally, by setting $\hat{\mathbf{y}}_i = \tilde{\mathbf{y}}^{(n)}$, we notice slightly worse results than our proposed re-labelling from Eq. (7.5).

7.3.4 Comparison with SOTA Methods

We compare our AsyCo with the following methods: 1) CE, which trains the classification network with standard CE loss on the noisy dataset; 2) Mixup [229], which employs mixup on the noisy dataset; 3) Forward [148], which estimates the noise transition matrix in a two-stage training pattern; 4) T-Revision [210], which finds reliable samples to replace anchor points for estimating transition matrix; 5) Reweight [120], which utilizes a class-dependent transition matrix to correct the loss function; 6) PTD-R-V [209], which proposes a part-dependent transition matrix for accurate estimation; 7) Decoupling [134], which trains two networks on samples whose predictions from the network are different; 8) Co-teaching [60], which trains two networks and select small-loss samples as clean samples; 9) MentorNet [82], which utilizes a teacher network for selecting noisy samples; 10) JoCoR [198], which trains two networks with joint regularization loss; 11) CausalNL [220], which discovers a causal relationship in noisy dataset and combines it with Co-Teaching; 12) CAL [241], which uses second-order statistics with a new loss function; 12) kMEIDTM [26], which learns instance-dependent transition matrix by applying manifold regularization during the training; 13) DivideMix [106], which combines semi-supervised learning, sample selection and Co-Teaching to achieve SOTA results; 14) FaMUS [216], which is a meta-learning method that learns the weight of training samples to improve the meta-learning update process; 15) Nested [25], which is a novel feature compression method that uses nested dropout to regularize features when training with noisy label–this approach can be combined with existing techniques such as Co-Teaching [60]; and 16) PLC [233], which is a method that produces soft pseudo label when learning with label noise.

Synthetic Noise Benchmarks. The experimental results of our proposed AsyCo with instance-dependent noise on CIFAR10/100 are shown in Tab. 7.3. We reproduce DivideMix [106] in this setup with single model at inference time denoted by $\theta^{(1)}$ and also the original ensemble inference. Compared with the best baselines, our method achieves large improvements for all noise rates. On CIFAR10, we achieve $\approx 1.5\%$ improvements for low noise rates and $\approx 1\%$ to 5% improvements for high noise rates. For CIFAR100, we improve between $\approx 1.5\%$ and $\approx 7\%$ for many noise rates. Note that our result is achieved without using small-loss sample selection, which is a fundamental technique for most noisy label learning methods [60, 82, 106]. The superior performance of AsyCo indicates that our multi-view consensus for sample selection and top-rank re-labelling are effective when learning with label noise.

Real-world Noisy-label Datasets. In Table 7.4, we present results on Red Mini-

Method	Noise rate			
	0.2	0.4	0.6	0.8
CE	47.36	42.70	37.30	29.76
Mixup [229]	49.10	46.40	40.58	33.58
DivideMix [106]	50.96	46.72	43.14	34.50
MentorMix [79]	51.02	47.14	43.80	33.46
FaMUS [216]	51.42	48.06	45.10	35.50
Ours	59.40	55.08	49.78	41.02
Ours w/ multi-label			52.48	42.76
Ours $2\times n_\theta$ test	61.98	57.46	51.86	42.58

Table 7.4: Test accuracy (%) of different methods on Red Mini-ImageNet with different noise rates. Baselines results are from FaMUS [216]. Best results with single/ensemble inferences are labelled with red/green.

Single	Methods	CE	Forward [148]	PTD-R-V [209]	ELR [119]	kMEIDTM [26]	Ours
	Accuracy	68.94	69.84	71.67	72.87	73.34	73.60
Ensemble	Methods	Co-Teaching [60]	Co-Teaching+ [224]	JoCoR [198]	CausalNL [220]	DivideMix [106]	Ours $2\times n_\theta$
	Accuracy	69.21	59.3	70.3	72.24	74.60	74.43

Table 7.5: Test accuracy (%) of different methods on Clothing1M. Best single/ensemble inference results are labelled with red/green.

Method	Accuracy
CE	79.4
Nested [25]	81.3
Dropout + CE [25]	81.1
SELFIE [171]	81.8
JoCoR [198]	82.8
PLC [233]	83.4
Nested + Co-Teaching [25]	84.1
Ours	85.6
Ours $2\times n_\theta$	86.3

Table 7.6: Test accuracy (%) of different methods on Animal-10N. Baselines results are presented with Nested Dropout [25]. Best single/ensemble inference results are labelled with red/green.

ImageNet [79]. Our method achieves SOTA results for all noise rates with 4% to 8% improvements in single model inference and 7% to 10% in ensemble inference. The improvement is significant compared with FaMUS [216] with a gap of more than 6%. Compared with DivideMix [106], our method achieves between 6% and 10% improvements. In Table 7.6, we present the results for Animal 10N [171], where the previous

SOTA method was Nested Dropout + Co-Teaching [25], which achieves 84.1% accuracy. Our method achieves 85.6% accuracy, which is 2.2% higher than previous SOTA. Additionally, our ensemble version achieves 86.34% accuracy, which improves 1% more compared to our single inference model, yielding a new SOTA result. In Table 7.5, we show our result on Clothing1M [211]. In the single model setup, our model outperforms all previous SOTA methods. In the ensemble inference setup, our model shows comparable performance with the SOTA method DivideMix [106] and outperforms all other methods. Compared with other methods based on prediction disagreement [60, 198, 224], our model improves by at least 3%. The performance on these three real-world datasets indicates the superiority of our proposed AsyCo.

7.4 Conclusion

In this work, we introduced a new noisy label learning method called AsyCo. Unlike previous SOTA noisy label learning methods that train two models with the same strategy and select small-loss samples, AsyCo explores two different training strategies and use multi-view consensus for sample selection. We show in experiments that AsyCo outperforms previous methods in both synthetic and real-world benchmarks. Our empirical analysis of AsyCo explores various subset selection strategies for sample selection and re-labelling, which show the importance of our design decisions. For future work, we will explore lighter models for the reference net as only rank prediction is required. We will also explores our method with out-of distribution (OOD) noise.

Statement of Authorship

Title of Paper	Generative Noisy-Label Learning by Implicit Dicriminative Approximation with Partial Label Prior
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted work written in manuscript style
Publication Details	

Principal Author

Name of Principal Author (Candidate)	Fengbei Liu
Contribution to the Paper	Proposed the ideas, conducted experiments and wrote the paper and revision
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: 09/14/2023

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 in 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 and revision
Signature	Date: 09/15/2023

Name of Co-Author	Chong Wang		
Contribution to the Paper	Discussion and commented on the revision		
Signature		Date	

Name of Co-Author	Yuyuan Liu		
Contribution to the Paper	Discussion and commented on the revision		
Signature		Date	

Name of Co-Author	Gustavo Carneiro		
Contribution to the Paper	Discussion and wrote the revision		
Signature		Date	22/09/2023

Chapter 8

Generative Noisy-label Learning by Implicit Discriminative Approximation with Partial Label Prior

Abstract

The learning with noisy labels has been addressed with both discriminative and generative models. Although discriminative models have dominated the field due to their simpler modeling and more efficient computational training processes, generative models offer a more effective means of disentangling clean and noisy labels and improving the estimation of the label transition matrix. However, generative approaches maximize the joint likelihood of noisy labels and data using a complex formulation that only indirectly optimizes the model of interest associating data and clean labels. Additionally, these approaches rely on generative models that are challenging to train and tend to use uninformative clean label priors. In this paper, we propose a new generative noisy-label learning approach that addresses these three issues. First, we propose a new model optimisation that directly associates data and clean labels. Second, the generative model is implicitly estimated using a discriminative model, eliminating the inefficient training of a generative model. Third, we propose a new informative label prior inspired by partial label learning as supervision signal for noisy label learning. Extensive experiments on several noisy-label benchmarks demonstrate that our generative model provides state-of-the-art results while maintaining a similar computational complexity as discriminative models.

8.1 Introduction

Deep neural network (DNN) has achieved remarkable success in computer vision [67, 97], natural language processing (NLP) [41, 223] and medical image analysis [113, 192]. However, DNNs often require massive amount of high-quality annotated data for supervised training [40], which is challenging and expensive to acquire. To alleviate such problem, some datasets have been annotated via crowdsourcing [211], from search engines [171], or with NLP from radiology reports [192]. Although these cheaper annotation processes enable the construction of large-scale datasets, they also introduce noisy labels for model training, resulting in performance degradation. Therefore, novel learning algorithms are required to robustly train DNN models when training sets contain noisy labels.

The main challenge in noisy-label learning is that we only observe the data, represented by random variable X, and respective noisy label, denoted by variable $\tilde{Y}$, but we want to estimate the model $p(Y|X)$, where Y is the hidden clean label variable. Most methods proposed in the field resort to two discriminative learning strategies: sample selection and noise transition matrix. *Sample selection* [1, 60, 106] optimises the model $p_\theta(Y|X)$, parameterised by θ, with maximum likelihood optimisation restricted to pseudo-clean training samples, as follows

$$\theta^* = \arg\max_\theta \mathbb{E}_{P(X,\tilde{Y})}\left[\mathsf{clean}(X,\tilde{Y}) \times p_\theta(\tilde{Y}|X)\right], \text{ where } \mathsf{clean}(X=\mathbf{x},\tilde{Y}=\tilde{\mathbf{y}}) = \begin{cases} 1, \text{ if } Y=\tilde{\mathbf{y}} \\ 0, \text{ otherwise} \end{cases}, \tag{8.1}$$

and $P(X,\tilde{Y})$ is the distribution used to generate the noisy-label and data points for the training set. Note that $\mathbb{E}_{P(X,\tilde{Y})}\left[\mathsf{clean}(X,\tilde{Y}) \times p_\theta(\tilde{Y}|X)\right] \equiv \mathbb{E}_{P(X,Y)}\left[p_\theta(Y|X)\right]$ if the function $\mathsf{clean}(.)$ successfully selects the clean-label training samples. Unfortunately, $\mathsf{clean}(.)$ usually relies on the *small-loss hypothesis* [3] for selecting $R\%$ of the smallest loss training samples, which offers little guarantees of successfully selecting clean-label samples. Approaches based on noise *transition matrix* [26, 148, 209] aim to estimate a clean-label classifier and a label transition, as follows:

$$\theta^* = \arg\max_\theta \mathbb{E}_{P(X,\tilde{Y})}\left[\sum_Y p(\tilde{Y},Y|X)\right] = \arg\max_{\theta_1,\theta_2} \mathbb{E}_{P(X,\tilde{Y})}\left[\sum_Y p_{\theta_1}(\tilde{Y}|Y,X)p_{\theta_2}(Y|X)\right], \tag{8.2}$$

where $\theta = [\theta_1, \theta_2]$, $p_{\theta_1}(\tilde{Y}|Y,X)$ represents a label-transition matrix, often simplified to be class-independent with $p_{\theta_1}(\tilde{Y}|Y) = p_{\theta_1}(\tilde{Y}|Y,X)$. Since we do not have access to the label transition matrix, we need to estimate it from the noisy-label training set, which is challenging because of identifiability issues [122], making necessary the use of anchor point [148] and regularisations [26].

On the other hand, generative learning models [5, 49, 220] assume a generative process for X and Y, as described in Fig. 8.1. These methods are trained to maximise the data likelihood $p(\tilde{Y},X) = \int_{Y,Z} p(X|Y,Z)p(\tilde{Y}|Y,X)p(Y)p(Z)dYdZ$, where Z denotes a

latent variable representing a low-dimensional representation of the image, and Y is the latent clean label. Y and Z are independent of each other and jointly generate X. This optimisation requires a variational distribution $q_\phi(Y, Z|X)$ to maximise the evidence lower bound (ELBO): with

$$\theta_1^*, \theta_2^*, \phi^* = \arg\max_{\theta_1,\theta_2,\phi} \mathbb{E}_{q_\phi(Y,Z|X)} \left[\log \left(p_{\theta_1}(X|Y,Z) p_{\theta_2}(\tilde{Y}|X,Y) p(Y) p(Z) / q_\phi(Y,Z|X) \right) \right], \tag{8.3}$$

where $p_{\theta_1}(X|Y,Z)$ denotes an image generative model, $p_{\theta_2}(\tilde{Y}|X,Y)$ represents the label transition model, $p(Z)$ is the latent image representation prior (commonly assumed to a standard normal distribution), and $p(Y)$ is the clean label prior (usually assumed to be a non-informative prior based on a uniform distribution). Such generative strategy is sensible because it disentangles the true and noisy labels and improves the estimation of the label transition model [220]. A limitation of the generative strategy is that it optimises $p(\tilde{Y}, X)$ instead of directly optimising $p(X|Y)$ or $p(Y|X)$. Also, compared with the discriminative strategy, the generative approach requires the generative model $p_{\theta_1}(X|Y,Z)$ that is challenging to train. This motivates us to ask the following question: **Can we directly optimise the generative goal $p(X|Y)$, with a similar computational cost as the discriminative strategy and accounting for an informative prior for the latent clean label Y?**

In this paper, we propose a new generative noisy-label learning method to directly optimise $p(X|Y)$ by maximising $\mathbb{E}_{q(Y|X)}[\log p(X|Y)]$ using a variational posterior distribution $q(Y|X)$. This objective function is decomposed into three terms: a label-transition model $\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}[\log p(\tilde{\mathbf{y}}|\mathbf{x}, \mathbf{y})]$, an image generative model $\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right]$, and a Kullback–Leibler (KL) divergence regularisation term. We implicitly estimate the image generative term with the discriminative model $q(Y|X)$, bypassing the need to train a generative model [157]. Moreover, our formulation allows the introduction of an instance-wise informative prior $p(Y)$ inspired by partial-label learning [181]. This prior is re-estimated at each training epoch to cover a small number of label candidates if the model is certain about the training label. Conversely, when the model is uncertain about the training label, then the label prior will cover a large number of label candidates, which also serve as a regularisation of noisy label training. Our formulation only requires a discriminative model and a label transition model, making it computationally less expensive than other generative approaches [5, 49, 220]. Overall, our contributions can be summarized as follows:

- We introduce a new generative framework to handle noisy-label learning by directly optimising $p(X|Y)$.

- Our generative model is implicitly estimated with a discriminative model, making it computationally more efficient than previous generative approaches [5, 49, 220].

- Our framework allows us to place an informative instance-wise prior $p(Y)$ for latent clean label Y. Inspired by partial label learning [188], $p(Y)$ is constructed to dynamically decrease uncertainty when the model has a large probability of high coverage, and increase uncertainty if the model has a low probability of high coverage.

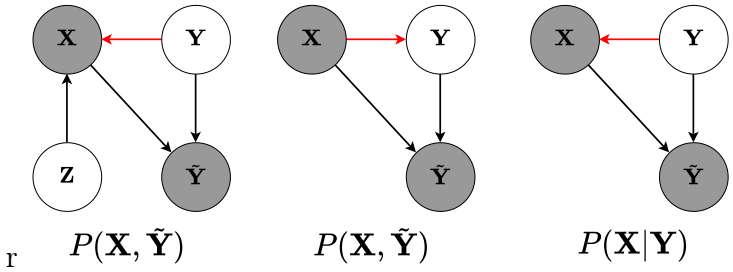
We conduct extensive experiments on both synthetic and real-world noisy-label benchmarks that show that our method provides state-of-the-art (SOTA) results and enjoy a similar computational complexity as discriminative approaches.

8.2 Related Work

Sample selection. The discriminative learning strategy based on sample selection from (8.1) needs to handle two problems: 1) the definition of $\mathsf{clean}(.)$, and 2) what to do with the samples classified as noisy. Most definitions of $\mathsf{clean}(.)$ resort to classify small-loss samples [3] as pseudo-clean [1, 20, 60, 81, 106, 134, 162, 198]. Other approaches select clean samples based on the K nearest neighbor classification in an intermediate deep learning feature spaces [145, 195], distance to the class-specific eigenvector from the gram matrix eigen-decomposition using intermediate deep learning feature spaces [91], uncertainty measures [95], or prediction consistency between teacher and student models [87]. After sample classification, some methods will discard the noisy-label samples for training [20, 81, 134, 162], while others use them for semi-supervised learning [106]. The main issue with this strategy is that it does not try to disentangle the clean and noisy-label from the samples.

Label transition model. The discriminative learning strategy based on the label transition model from (8.2) depends on a reliable estimation of $p(\tilde{Y}|Y, X)$ [26, 148, 209]. Forward-T [148] uses an additional classifier and anchor points from clean-label samples to learn a class-dependent transition matrix. Part-T [209] estimates an instance-dependent model. MEDITM [26] uses manifold regularization for estimating the label-transition matrix. In general, the estimation of this label transition matrix is under-constrained, leading to the identifiability problem [122], which is addressed with the formulation of strong assumptions [148], or the use of additional labels per training sample [122].

Generative modelling. Generative modeling for noisy-label learning [5, 49, 220] explores different graphical models (see Fig. 8.1) to enable the estimation of clean labels per image. Specifically, CausalNL [220] and InstanceGM [49] assume that the latent clean label Y causes X, and the noisy label $\tilde{Y}$ is generated from X and Y. Alternatively, NPC [5] assumes that X causes Y and proposes a post-processing calibration for noisy label learning. One drawback of generative modeling is that instead of directly optimising the models of interest $p(X|Y)$ or $p(Y|X)$, it optimises the joint distribution of visible variables $p(X, \tilde{Y})$. Even though maximising the likelihood of the



r

Figure 8.1: Generative noisy-label learning models and their corresponding optimisation goal, where the red arrow indicates the different causal relationships between X and Y. Left is CausalNL/InstanceGM [49, 220], middle is NPC [5] and right is ours.

visible data is sensible, it only produces the models of interest as a by-product of the process. Furthermore, these methods require the computationally complex training of a generative model, and usually rely on non-informative label priors.

Clean label prior. Our clean-label prior $p(Y)$ constrains the clean label to a set of label candidates for a particular training sample. Such label candidates change during training following two design principles: 1) increase clean label coverage, and 2) reduce the uncertainty of the label prior. The increase of coverage improves the chances of including the correct clean label into the prior. Given that this decreases the quality of the supervisory training signal, the second design principle regularises the training by reducing the number of label candidates in $p(Y)$. Such dynamic prior distribution may resemble Mixup [229], label smoothing [129] or re-labeling [106] techniques that are commonly used in label noise learning. However, these approaches do not simultaneously follow the two design principles mentioned above. Mixup [229] and label smoothing [129] are effective approaches for designing soft labels for noisy label learning, but both aim to increase coverage, disregarding label uncertainty. Re-labeling switches the supervisory training signal to a more likely pseudo label, so it is very efficient, but it has limited coverage.

Partial label learning In partial label learning (PLL), each image is associated with a candidate label set defined as a partial label [181]. The goal of PLL is to predict the single true label associated with each training sample, assuming that the ground truth label is one of the labels in its candidate set. PICO [188] uses contrastive learning in an EM optimisation to address PLL. CAV [228] proposes class activation mapping to identify the true label within the candidate set. PRODEN [130] progressively identifies the true labels from a candidate set and updates the model parameter. The design of our informative clean label prior $p(Y)$ is inspired from PLL, but unlike PLL, there is no guarantee that the multiple label candidates in our prior contain the true label. Furthermore, the size of our candidate label set is determined by the probability that the training sample label is clean, where a low probability induces a prior with a large

number of candidates, while a high probability induces a prior with a small number of label candidates.

8.3 Method

We denote the noisy training set as $\mathcal{D} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^{H \times W \times C}$ is the input image of size $H \times W$ with C colour channels, $\tilde{\mathbf{y}}_i \in \mathcal{Y} \subset \{0,1\}^{|\mathcal{Y}|}$ is the observed noisy label. We also have $\mathbf{y}$ as the unobserved clean label. We formulate our model with generative model that starts with the sampling of a label $\mathbf{y} \sim p(Y)$. This is followed by the clean-label conditioned generation of an image with $\mathbf{x} \sim p(X|Y = \mathbf{y})$, which are then used to produce the noisy label $\tilde{\mathbf{y}} \sim p(\tilde{Y}|Y = \mathbf{y}, X = \mathbf{x})$ (hereafter, we omit the variable names to simplify the notation). Below, in Sec. 8.3.1, we introduce our model and the optimisation goal. In Sec. 8.3.2 we describe how to construct informative prior, and the overall training algorithm is presented in Sec. 8.3.3.

8.3.1 Model

We aim to optimize the generative model $\log p(\mathbf{x}|\mathbf{y})$, which can be decomposed as follows:

$$\log p(\mathbf{x}|\mathbf{y}) = \log \frac{p(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{x})}{p(\tilde{\mathbf{y}}|\mathbf{x}, \mathbf{y})p(\mathbf{y})}. \tag{8.4}$$

In (8.4), $p(\mathbf{y})$ represents the prior distribution of the latent clean label. The optimisation of $p(\mathbf{x}|\mathbf{y})$ can be achieved by introducing a variational posterior distribution $q(\mathbf{y}|\mathbf{x})$, with:

$$\begin{aligned} \log p(\mathbf{x}|\mathbf{y}) &= \log \frac{p(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{x})}{q(\mathbf{y}|\mathbf{x})} + \log \frac{q(\mathbf{y}|\mathbf{x})}{p(\tilde{\mathbf{y}}|\mathbf{x}, \mathbf{y})p(\mathbf{y})}, \\ \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log p(\mathbf{x}|\mathbf{y})\right] &= \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{x})}{q(\mathbf{y}|\mathbf{x})}\right] + \mathsf{KL}\Big[q(\mathbf{y}|\mathbf{x})||p(\tilde{\mathbf{y}}|\mathbf{x}, \mathbf{y})p(\mathbf{y})\Big], \end{aligned} \tag{8.5}$$

where $\mathsf{KL}[.]$ denotes the KL divergence, and

$$\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{x})}{q(\mathbf{y}|\mathbf{x})}\right] = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log p(\tilde{\mathbf{y}}|\mathbf{x}, \mathbf{y})\right] + \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right]. \tag{8.6}$$

Based on Eq. (8.5) and (8.6), the expected log likelihood of $p(\mathbf{x}|\mathbf{y})$ is defined as

$$\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log p(\mathbf{x}|\mathbf{y})\right] = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log p(\tilde{\mathbf{y}}|\mathbf{x}, \mathbf{y})\right] - \mathsf{KL}\left[q(\mathbf{y}|\mathbf{x})\|p(\mathbf{x}|\mathbf{y})p(\mathbf{y})\right] + \mathsf{KL}\left[q(\mathbf{y}|\mathbf{x})\|p(\tilde{\mathbf{y}}|\mathbf{x}, \mathbf{y})p(\mathbf{y})\right]. \tag{8.7}$$

In Eq. (8.7), we parameterise $q(\mathbf{y}|\mathbf{x})$ and $p(\tilde{\mathbf{y}}|\mathbf{x},\mathbf{y})$ with neural networks, as depicted in Figure 8.2. the generative term $p(\mathbf{x}|\mathbf{y})$ remains challenging to estimate due to its intractability. This is in part due to the infinite number of samples X that can be generated from their clean labels Y. One solution to mitigate such intractability is the use of a latent image representation Z to "anchor" the image generation process, as what previous works does [49, 220]. However, note that such image generation ability is in fact irrelevant for discriminative classification tasks that we aim to solve [5]. This additional modeling Z becomes troublesome with large resolution input and leads to sub-optimal reconstruction, which is a common issue for generative model (VAE). Furthermore, Z and Y jointly generate X, which imply that they need to be disentangled for classification task and such disentanglement is not the main goal of noisy label classification.

we assume that Z is unnecessary and $p(\mathbf{x}|\mathbf{y})$ is defined only on the finite number of training samples given by classification task. This assumption facilitate the direct optimisation of $p(\mathbf{x}|\mathbf{y})$ and alleviate the problematic training of an image generator [157]. This optimum is achieved by :

$$p(\mathbf{x}|\mathbf{y}) \propto \frac{q(\mathbf{y}|\mathbf{x})}{\sum_{i=1}^{|\mathcal{D}|} q(\mathbf{y}|\mathbf{x}_i)}. \tag{8.8}$$

Hence, the generative conditional $p(\mathbf{x}|\mathbf{y})$ is approximated only with finite number of samples of $\mathbf{x}$ given the latet labels in $\mathbf{y}$, making this term tractable. As mentioned in [157], the probabilities $p(\mathbf{x}|\mathbf{y})$ are larger for the data samples $\mathbf{x}$ for which $q(\mathbf{y}|\mathbf{x})$ is also large relative to the assignment to class $\mathbf{y}$ to all training samples.

8.3.2 Informative prior based on partial label learning

In Eq. (8.7), the clean label prior $p(\mathbf{y})$ is required. As mentioned in Sec. 8.2, we formulate $p(\mathbf{y})$ inspired from PLL [130, 188, 228]. However, it is worth noting that PLL has the partial label information available from the training set, while we have to dynamically build it during training. Therefore, the clean label prior $p(\mathbf{y})$ for each training sample is designed so that the hidden clean label has a high probability of being selected during most of the training. On one hand, we aim to have as many label candidates as possible during the training to increase the chances that $p(\mathbf{y})$ has a non-zero probability for the latent clean label. On the other hand, including all labels as candidates is a trivial solution that does not represent a meaningful clean label prior. These two seemingly contradictory goals target the maximisation of label coverage and minimisation of label uncertainty, defined by:

$$\text{Coverage} = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\sum_{j=1}^{|\mathcal{Y}|} \mathbb{1}\left(\mathbf{y}_i(j) \times p_i(j) > 0\right), \text{ and Uncertainty} = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\sum_{j=1}^{|\mathcal{Y}|} \mathbb{1}\left(p_i(j) > 0\right), \tag{8.9}$$

where $\mathbb{1}(.)$ is the indicator function. In (8.9), coverage increases by approximating $p(Y)$ to a uniform distribution, but uncertainty is minimised when the clean label $\mathbf{y}_i$ is assigned maximum probability. In general, training samples for which the model is certain about the clean label, should have $p(\mathbf{y}_i) = 1$, while training samples for which the model is uncertain about the clean label, should have $p(\mathbf{y}_i) < 1$ with other candidate labels with probability > 0. Therefore, the clean label prior is defined by:

$$p_i(j) = \frac{\tilde{\mathbf{y}}_i(j) + \mathbf{c}_i(j) + \mathbf{u}_i(j)}{Z}, \tag{8.10}$$

where Z is a normalisation factor to make $\sum_{j=1}^{|\mathcal{Y}|} p_i(j) = 1$, $\tilde{\mathbf{y}}_i$ is the noisy label in the training set, $\mathbf{c}_i$ denotes the label to increase coverage, and $\mathbf{u}_i$ represents the label to increase uncertainty, both defined below. Motivated by the early learning phenomenon [119], where clean labels tend to be fit earlier in the training than the noisy labels, we maximise coverage by sampling from a moving average of model prediction for each training sample $\mathbf{x}_i$ at iteration t with:

$$\mathcal{C}_i^{(t)} = \beta \times \mathcal{C}_i^{(t-1)} + (1 - \beta) \times \bar{\mathbf{y}}_i^{(t)}, \tag{8.11}$$

where $\beta \in [0, 1]$ and $\bar{\mathbf{y}}^{(t)}$ is the softmax output from the model that predicts the clean label from the data input $\mathbf{x}_i$. For Eq. (8.11), $\mathcal{C}_i^{(t)}$ denotes the categorical distribution of the most likely labels for the i^{th} training sample, which can be used to sample the one-hot label $\mathbf{c}_i \sim \mathsf{Cat}(\mathcal{C}_i^{(t)})$. The minimisation of uncertainty depends on our ability to detect clean-label and noisy-label samples. For clean samples, $p(\mathbf{y}_i)$ should converge to a one-hot distribution, maintaining the label prior focused on few candidate labels. For noisy samples, $p(\mathbf{y}_i)$ should be close to a uniform distribution to keep a large coverage of candidate labels. To compute the probability $w_i \in [0, 1]$ that a sample contains clean label, we use the sample selection approaches based on the unsupervised classification of loss values [106]. Then the label $\mathbf{u}_i$ is obtained by sampling from a uniform distribution of all possible labels proportionally to its probability of representing a noisy-label sample, with

$$\mathbf{u}_i \sim \mathcal{U}\left(\mathcal{Y}, \mathsf{round}(|\mathcal{Y}| \times (1 - w_i))\right), \tag{8.12}$$

where $\mathsf{round}(|\mathcal{Y}| \times (1 - w_i))$ represents the number of samples to be drawn from the uniform distribution rounded up to the closest integer.

8.3.3 Training

We can now return to the optimisation of Eq. (8.7), where we define the neural networks $g_\theta : \mathcal{X} \to \Delta^{|\mathcal{Y}|-1}$ that outputs the categorical distribution for the clean label in the probability simplex space $\Delta^{|\mathcal{Y}|-1}$ given an image $\mathbf{x} \in \mathcal{X}$, and $f_\phi : \mathcal{X} \times \Delta^{|\mathcal{Y}|-1} \to \Delta^{|\mathcal{Y}|-1}$ that outputs the categorical distribution for the noisy training label given an image

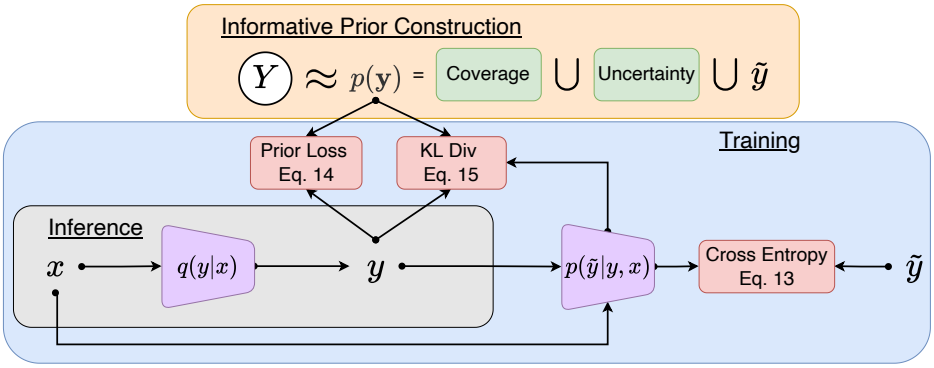


Figure 8.2: Training pipeline of our method. Shaded variables $\mathbf{x}$ and $\tilde{\mathbf{y}}$ are visible, and unshaded variable $\mathbf{y}$ is latent. We build $p(\mathbf{y})$ to represent candidate labels to approximate $\mathbf{y}$.

and the clean label distribution from $g_\theta(.)$. The first term in the right-hand side (RHS) in Eq. (8.7) is optimised with the cross-entropy loss:

$$\mathcal{L}_{CE}(\theta, \phi, \mathcal{D}) = \frac{1}{|\mathcal{D}| \times K} \sum_{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \mathcal{D}} \sum_{j=1}^{K} \ell_{CE}(\tilde{\mathbf{y}}_i, f_\phi(\mathbf{x}_i, \hat{\mathbf{y}}_{i,j})). \tag{8.13}$$

where $\{\hat{\mathbf{y}}_{i,j}\}_{j=1}^{K} \sim \mathsf{Cat}(g_\theta(\mathbf{x}_i))$, with $\mathsf{Cat}(.)$ denoting a categorical distribution. The second term in the RHS in Eq. (8.7) uses the estimation of $p(\mathbf{x}|\mathbf{y})$ from Eq. (8.8) to optimise the KL divergence:

$$\mathcal{L}_{PRI}(\theta, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \mathcal{D}} \mathsf{KL}\left[g_\theta(\mathbf{x}_i) \middle\| c_i \times \frac{g_\theta(\mathbf{x}_i)}{\sum_j g_\theta(\mathbf{x}_j)} \odot \mathbf{p}_i \right], \tag{8.14}$$

where $\mathbf{p}_i = [p_i(j = 1), ..., p_i(j = |\mathcal{Y}|)] \in \Delta^{|\mathcal{Y}|-1}$ is the clean label prior defined in Eq. (8.10), c_i is a normalisation factor, and $\odot$ is the element-wise multiplication. The last term in the RHS of Eq. (8.7) is the KL divergence between $q(\mathbf{y}|\mathbf{x})$ and $p(\tilde{\mathbf{y}}|\mathbf{x}, \mathbf{y})p(\mathbf{y})$, which represents the gap between $\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{y})]$ and $\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{x})}{q(\mathbf{y}|\mathbf{x})}\right]$. According to the expectation-maximisation (EM) derivation [39, 92], the smaller this gap, the better $q(\mathbf{y}|\mathbf{x})$ approximates the true posterior $p(\mathbf{y}|\mathbf{x})$, so the loss function associated with this third term is:

$$\mathcal{L}_{KL}(\theta, \phi, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \mathcal{D}} \mathsf{KL}\left[g_\theta(\mathbf{x}_i) \middle\| f_\phi(\mathbf{x}_i, g_\theta(\mathbf{x}_i)) \odot \mathbf{p}_i \right]. \tag{8.15}$$

Our final loss to minimise is

$$\mathcal{L}(\theta, \phi, \mathcal{D}) = \mathcal{L}_{CE}(\theta, \phi, \mathcal{D}) + \mathcal{L}_{PRI}(\theta, \mathcal{D}) + \mathcal{L}_{KL}(\theta, \phi, \mathcal{D}). \tag{8.16}$$

After training, a test image $\mathbf{x}$ is associated with a class with $g_\theta(\mathbf{x})$. An interesting point about this derivation is that the implicit approximation of $p(\mathbf{x}|\mathbf{y})$ enables the minimisation of the loss in (8.16) using regular stochastic gradient descent instead the computationally more complex expectation-maximisation (EM) algorithm [157].

8.4 Experiments

We show experimental results on instance-dependent synthetic and real-world label noise benchmarks with datasets CIFAR10/100 [96]. We also test on three instance-dependent real-world label noise datasets, namely: Animal-10N [171], Red Mini-ImageNet [81], and Clothing1M [211].

8.4.1 Datasets

CIFAR10/100 [96] contain a training set with 50K images and a testing of 10K images of size $32 \times 32 \times 3$, where CIFAR10 has 10 classes and CIFAR100 has 100 classes. We follow previous works [209] and synthetically generate instance-dependent noise (IDN) with rates in {0.2, 0.3, 0.4 ,0.5}. **CIFAR10N/CIFAR100N** is proposed by [203] to study real-world annotations for the original CIFAR10/100 images and we test our framework on {aggre, random1, random2, random3, worse} types of noise on CIFAR10N and {noisy} on CIFAR100N. **Red Mini-ImageNet** is a real-world dataset [81] with images annotated with the Google Cloud Data Labelling Service. This dataset has 100 classes, each containing 600 images from ImageNet, where images are resized to 32×32 pixels from the original 84×84 to enable a fair comparison with other baselines [216]. **Animal 10N** [171] is a real-world dataset containing 10 animal species with five pairs of similar appearances (wolf and coyote, hamster and guinea pig, etc.). The training set size is 50K and testing size is 10K, where we follow the same set up as [25]. **Clothing1M** is a real-world dataset with 100K images and 14 classes. The labels are automatically generated from surrounding text with an estimated noise ratio of 38.5%. The dataset also contains clean training, clean validation and clean test sets with 50K, 14K and 10K images, respectively, but we do not use the clean training and validation sets. The clean testing is only used for measuring model performance.

8.4.2 Practical considerations

We follow commonly used experiment setups, including network architecture, hyper-parameter setups for all benchmarks and describe more details in Appendix E. For the hyper-parameter setup, K in (8.13) is set to 1, and β in Eq. (8.11) is set to 0.9. For w in Eq. (8.12), we follow the commonly used Gaussian Mixture Model (GMM) unsupervised classification from [106]. For warmup epochs, w is randomly generated

Method	CIFAR10			
	20%	30%	40%	50%
CE	86.93±0.17	82.42±0.44	76.68±0.23	58.93± 1.54
DMI [215]	89.99± 0.15	86.87± 0.34	80.74± 0.44	63.92±3.92
Forward [148]	89.62±0.14	86.93±0.15	80.29±0.27	65.91±1.22
CoTeaching [60]	88.43±0.08	86.40±0.41	80.85±0.97	62.63± 1.51
TMDNN [218]	88.14± 0.66	84.55±0.48	79.71±0.95	63.33± 2.75
PartT [209]	89.33± 0.70	85.33±1.86	80.59±0.41	64.58± 2.86
kMEIDTM [26]	92.26± 0.25	90.73± 0.34	85.94± 0.92	73.77±0.82
CausalNL [220]	81.47± 0.32	80.38± 0.44	77.53± 0.45	67.39±1.24
Ours	**92.65±0.13**	**91.96±0.20**	**91.02±0.44**	**89.94±0.45**

Table 8.1: Accuracy (%) on the test set for IDN problems on CIFAR10. Most results are from [26]. Experiments are repeated 3 times to compute mean±standard deviation. Top part shows discriminative and bottom shows generative models. Best results are highlighted.

from a uniform distribution. Note that the approximation of the generative model from (8.8) is done within each batch, not the entire the dataset. Also, following the discussion by Rolf et al. [157], the minimisation of $\mathcal{L}_{PRI}(.)$ can be done with the reversed KL using $\mathsf{KL}\left[c_i \times \frac{g_\theta(\mathbf{x}_i)}{\sum_j g_\theta(\mathbf{x}_j)} \odot \mathbf{p}_i \middle\| g_\theta(\mathbf{x}_i)\right]$. This reversed KL divergence also provides solutions where the model and implied posterior are close. In fact, the KL and reversed KL losses are equivalent when $\sum_j g_\theta(\mathbf{x}_j)$ has a uniform distribution over the classes in $\mathcal{Y}$ and the prior $\mathbf{p}_i$ is uniform in the negative labels. We tried the optimisation using both versions of the KL divergence (i.e., the one in (8.14) and the one above in this section), with the reversed one generally producing better results, as shown in the ablation study in Sec. 8.4.4. For all experiments in Sec. 8.4.3, we rely on the reversed KL loss. For the real-world datasets Animal-10N, Red Mini-ImageNet and Clothing1M we also test our model with the training and testing of an ensemble of two networks. Our code is implemented in Pytorch and experiments are performed on RTX 3090.

8.4.3 Experimental Results

Synthetic benchmarks. The experimental results of our method with IDN problems on CIFAR10/100 are shown in Tab.8.1 and Tab.8.2. Compared with the previous SOTA kMEDITM [26], on CIFAR10, we achieve competitive performance on low noise rates and up to 16% improvements for high noise rates. For CIFAR100, we consistently improve 2% to 4% in all noise rates. Compared with the previous SOTA generative model CausalNL [220], our improvement is significant for all noise rates. The superior performance of our method indicates that our implicit generative modelling and clean

Method	CIFAR100			
	20%	30%	40%	50%
CE	63.94±0.51	61.97±1.16	58.70±0.56	56.63±0.69
DMI [215]	64.72±0.64	62.8±1.46	60.24±0.63	56.52±1.18
Forward [148]	67.23±0.29	65.42±0.63	62.18±0.26	58.61±0.44
CoTeaching [60]	67.40±0.44	64.13±0.43	59.98±0.28	57.48±0.74
TMDNN [218]	66.62±0.85	64.72±0.64	59.38±0.65	55.68±1.43
PartT [209]	65.33±0.59	64.56±1.55	59.73±0.76	56.80±1.32
kMEIDTM [26]	69.16±0.16	66.76±0.30	63.46±0.48	59.18±0.16
CausalNL [220]	41.47±0.43	40.98±0.62	34.02±0.95	32.13±2.23
Ours	**71.24±0.43**	**69.64±0.78**	**67.48±0.85**	**63.60±0.17**

Table 8.2: Accuracy (%) on the test set for IDN problems on CIFAR100. Most results are from [26]. Experiments are repeated 3 times to compute mean±standard deviation. Top part shows discriminative and bottom shows generative models. Best results are highlighted.

Method	CIFAR10N					CIFAR100N
	Aggregate	Random 1	Random 2	Random 3	Worst	Noisy
CE	87.77±0.38	85.02±0.65	86.46±1.79	85.16±0.61	77.69±1.55	55.50±0.66
Forward T [148]	88.24±0.22	86.88±0.50	86.14±0.24	87.04±0.35	79.79±0.46	57.01±1.03
T-Revision [210]	88.52±0.17	88.33±0.32	87.71±1.02	80.48±1.20	80.48±1.20	51.55±0.31
Positive-LS [129]	91.57±0.07	89.80±0.28	89.35±0.33	89.82±0.14	82.76±0.53	55.84±0.48
F-Div [201]	91.64±0.34	89.70±0.40	89.79±0.12	89.55±0.49	82.53±0.52	57.10±0.65
Negative-LS [199]	91.97±0.46	90.29±0.32	90.37±0.12	90.13±0.19	82.99±0.36	58.59±0.98
CORES^2 [27]	91.23±0.11	89.66±0.32	89.91±0.45	89.79±0.50	83.60±0.53	**61.15±0.73**
VolMinNet [110]	89.70±0.21	88.30±0.12	88.27±0.09	88.19±0.41	80.53±0.20	57.80±0.31
CAL [241]	91.97±0.32	90.93±0.31	90.75±0.30	90.74±0.24	85.36±0.16	**61.73±0.42**
Ours	**92.57±0.20**	**91.97±0.09**	**91.42±0.06**	**91.83±0.12**	**86.99±0.36**	**61.54±0.22**

Table 8.3: Accuracy (%) on the test set for CIFAR10N/100N. Results are taken from [203] using methods containing a single classifier with ResNet-34. Best results are highlighted.

label prior construction is effective when learning with label noise.

Real-world benchmarks. In Tab.8.3, we show the performance of our method on the CIFAR10N/100N benchmark. Compared with other single-model baselines, our method achieves at least 1% improvement on all noise rates on CIFAR10N, and it has a competitive performance on CIFAR100N. The Red Mini-ImageNet results in Tab.8.4 (left) show that our method achieves SOTA results for all noise rates with 2% improvements using a single model and 6% improvements using the ensemble of two models. The improvement is substantial compared with previous SOTA FaMUS [216] and DivideMix [106]. In Tab.8.4(right), our single-model result on Animal-10N achieves 1% improvement with respect to the single-model SELFIE [171]. Considering our

Method	Noise rate			
	0.2	0.4	0.6	0.8
CE	47.36	42.70	37.30	29.76
Mixup [229]	49.10	46.40	40.58	33.58
DivideMix [106]	50.96	46.72	43.14	34.50
MentorMix [79]	51.02	47.14	43.80	33.46
FaMUS [216]	51.42	48.06	45.10	35.50
Ours	53.34	49.56	44.08	36.70
Ours (ensemble)	**57.56**	**52.68**	**47.12**	**39.54**

Method	Accuracy
CE	79.4
SELFIE [171]	81.8
JoCoR [198]	82.8
PLC [233]	83.4
Nested + Co-T [25]	84.1
InstanceGM [49]	84.6
Ours	82.7
Ours (ensemble)	**85.7**

Table 8.4: Test accuracy (%) on Red Mini-ImageNet (Left) with different noise rates and baselines from FaMUS [216], and on Animal-10N (Right), with baselines from [25]. Best results are highlighted.

approach with an ensemble of two models, we achieve a 1% improvement over the SOTA Nested+Co-teaching [25]. Our ensemble-model result on Clothing1M in Tab.8.5 shows a competitive performance of 74.4%, which is 2% better than the previous SOTA generative model CausalNL [220].

8.4.4 Analysis

Ablation The ablation analysis of our method is shown in Tab.8.6 with the IDN problems on CIFAR10. First row ($\mathcal{L}_{CE}$) shows the results of the training with a cross-entropy loss using the training samples and labels in $\mathcal{D}$. The second row ($\mathcal{L}_{CE}$ + $\mathcal{L}_{CE\_PRI}$ +$\mathcal{L}_{KL}$) shows the result of our method, replacing the KL divergence in $\mathcal{L}_{PRI}$ as defined in (8.14), by a cross entropy loss, as suggested in [157]. Next, the third row ($\mathcal{L}_{CE}$ + $\mathcal{L}_{PRI}$ + $\mathcal{L}_{KL}$) shows our method with the loss defined in (8.16). As mentioned in Sec. 8.4.2, following [157], we tested the reverse KL divergence in $\mathcal{L}_{PRI}$ from (8.14), first, in the fourth row ($\mathcal{L}_{CE}$ + $\mathcal{L}_{PRI}$ reversed) by optimising the lower bound to $\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{y})]$, and then by optimising the whole objective function from (8.16) in the last row ($\mathcal{L}_{CE}$ + $\mathcal{L}_{PRI}$ reversed + $\mathcal{L}_{KL}$ (Ours)). In general, notice that the reversed $\mathcal{L}_{PRI}$ improves the results; the KL divergence in $\mathcal{L}_{PRI}$ works better than the CE loss; and the optimisation of the whole loss in (8.16) is better than optimising the lower bound, which justifies the inclusion of $\mathcal{L}_{KL}(.)$ in the loss.

Coverage and uncertainty visualisation We visualise coverage and uncertainty from Eq. (8.9) at each training epoch for IDN CIFAR10/100 and CIFAR10N setups. In all cases, label coverage increases as training progresses, indicating that our prior tends to always cover the clean label. In fact, coverage reaches nearly 100% for CIFAR10 at 20% IDN and 97% for 50% IDN. Furthermore, for CIFAR100 at 50% IDN, we achieve 82% coverage, and for CIFAR10N "worse", we reach 92% coverage. In terms of uncertainty, we notice a steady reduction as training progresses for all problems,

CE	Forward [148]	PTD-R-V [209]	ELR [119]	kMEIDTM [26]	CausalNL [220]	Ours (ensemble)
68.94	69.84	71.67	72.87	73.34	72.24	**74.35**

Table 8.5: Test accuracy (%) on the test set of Clothing1M. Results are obtained from their respective papers. We only use the noisy training set for training. Best results are highlighted.

Method	CIFAR10			
	20%	30%	40%	50%
$\mathcal{L}_{CE}$	86.93	82.42	76.68	58.93
$\mathcal{L}_{CE}$ + $\mathcal{L}_{CE\_PRI}$ +$\mathcal{L}_{KL}$	85.96	82.74	78.34	73.72
$\mathcal{L}_{CE}$ + $\mathcal{L}_{PRI}$ + $\mathcal{L}_{KL}$	91.36	90.88	90.25	88.47
$\mathcal{L}_{CE}$ + $\mathcal{L}_{PRI}$ reversed	92.40	90.23	87.75	80.46
$\mathcal{L}_{CE}$ + $\mathcal{L}_{PRI}$ reversed + $\mathcal{L}_{KL}$ (Ours)	92.65	91.96	91.02	89.94

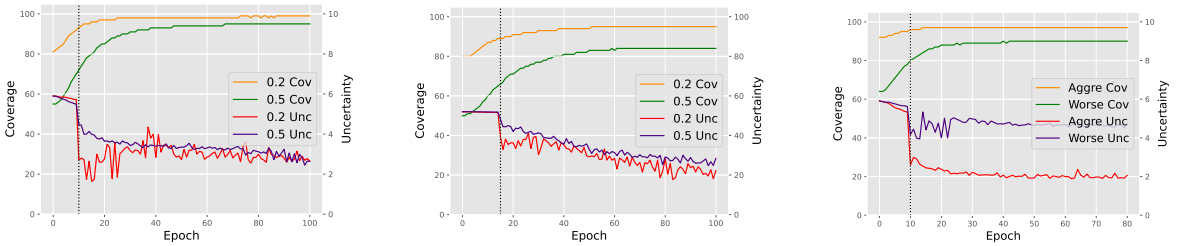
Table 8.6: Ablation analysis of our proposed method. Please see text for details.

	CE	DivideMix [106]	CausalNL [220]	InstanceGM [49]	Ours
CIFAR	2.1h	7.1h	3.3h	30.5h	2.3h
Clothing1M	4h	14h	10h	43h	4.5h

Table 8.7: Running times of various methods on CIFAR100 with 50% IDN and Clothing1M using the hardware listed in Sec. 8.4.2.

where the uncertainty values tend to be slightly higher for the problems with higher noise rates and more classes. For instance, uncertainty is between 2 and 3 for the for CIFAR10's IDN benchmarks, increasing to be between 2 and 4 for CIFAR10N. For CIFAR100's IDN benchmarks, uncertainty is between 20 and 30. These results suggest that our prior clean label distribution is effective at selecting the correct clean label while reducing the number of label candidates during training.

Training time comparison One of the advantages of our approach is its efficient training algorithm, particularly when compared with other generative and discriminative methods. Tab. 8.7 shows the training time for competing approaches on CIFAR100 with 50% IDN and Clothing1M using the hardware specified in Sec. 8.4.2. In general, our method has a smaller training time than competing approaches, being equivalent to the training with CE loss, around 2× faster than CausalNL [220], 3× faster than DivideMix [106], and 10× faster than InstanceGM [49].



(a) CIFAR10-IDN at 20% and 50% noise rates

(b) CIFAR100-IDN at 20% and 50% noise rates

(c) CIFAR10N "Worse" and "Aggre" noise rates

Figure 8.3: Coverage (Cov) and uncertainty (Unc) for (a) CIFAR10-IDN (20% and 50%), (b) CIFAR100-IDN (20% and 50%), and (c) CIFAR10N ("Worse" and "Aggre"). Y-axis shows coverage (left) and uncertainty (right). The dotted vertical line indicates the end of warmup training.

8.5 Conclusion

In this paper, we presented a new learning algorithm to optimise a generative model represented by $p(X|Y)$ that directly associates data and clean labels instead of maximising the joint data likelihood, denoted by $p(X, \tilde{Y})$. Our optimisation implicitly estimates $p(X|Y)$ with the discriminative model $q(Y|X)$ [157] eliminating the inefficient generative model training. Furthermore, we introduce an informative clean label prior, inspired by partial-label learning [181], to cover a small number of label candidates when the model is certain about the training label, and to cover a large number of label candidates, otherwise. Results on synthetic and real-world noisy-label benchmarks show that our generative method has SOTA results, but with complexity comparable to discriminative models.

A limitation of the proposed method that needs further exploration is a comprehensive study of the model for $q(Y|X)$. In fact, the results shown in this paper are surprisingly competitive given that we use fairly standard models for $q(Y|X)$ without exploring sophisticated noisy-label learning techniques. In the future, we will use more powerful models for $q(Y|X)$. Another issue of our model is the difficulty to estimate $p(X|Y)$ in real-world datasets containing images of high resolution. We will study more adequate ways to approximate $p(X|Y)$ in such scenario using data augmentation strategies to increase the size of the dataset and using a more effective method to approximate the calculation of $p(X|Y)$ for the whole training set instead of only the mini-batch dataset.

Chapter 9

Conclusion and Discussion

In this thesis, we have devised efficient weakly-supervised methodologies for tasks related to computer vision and medical image analysis. Initially, we examined the challenges associated with current semi-supervised learning (SSL) techniques, such as severe class imbalance and multi-label issues arising from the infrequency of diseases and the complexities in determining the optimal augmentation strategy from computer vision to medical image analysis.

Historically, methods have predominantly utilized ImageNet [40] pre-trained weights, which are not entirely compatible with medical image analysis tasks. Motivated by recent advancements in self-supervised learning, which focuses on learning task-relevant feature representation, we introduced a novel SSL approach named Self-Supervised Mean-Teacher for Semi-Supervised (S2MTS2) learning. This method combines self-supervised pre-training and mean-teacher fine-tuning. The primary innovation of our approach is the application of joint contrastive learning to pre-train a student-teacher model using unlabeled data. This enhances the feature representation and ensures consistency between the student and teacher models. We evaluated our method using three datasets: ChestX-ray14 [192], CheXpert [75], and ISIC2018 [182], demonstrating that it significantly outperforms existing state-of-the-art SSL methods. Additionally, we conducted an ablation study to assess the impact of each component of our method. The results underscore the effectiveness and robustness of our approach for medical image classification under limited supervision.

We subsequently analyzed the pseudo-labelling approach for medical image analysis in semi-supervised learning. The challenge with the pseudo-labelling approach lies in designing an appropriate threshold for selecting high-confidence pseudo labels. This is particularly complex in multi-label classification, where defining a class-wise threshold without prior knowledge about class distribution or class correlation is difficult. To address this, we proposed a novel semi-supervised learning method for medical image analysis, termed Anti-Curriculum Pseudo-Labelling (ACPL). This method selects informative unlabeled samples and assigns them pseudo-labels based on a combination

of model prediction and K-Nearest Neighbors (KNN) voting. The primary innovation of our approach is the use of an anti-curriculum strategy to select samples with high uncertainty and diversity, and a label purification module to update the anchor set. We evaluated our method using two datasets: Chest X-ray14 and ISIC2018, demonstrating that it significantly outperforms existing state-of-the-art semi-supervised learning methods. Additionally, we conducted an ablation study to assess the impact of each component of our method. The results underscore the effectiveness and robustness of our approach for multi-label semi-supervised learning medical image classification under class imbalance.

Our methods for tackling semi-supervised learning in medical imaging lead to two different perspectives. S2MTS2 uses a large number of unlabelled samples to get domain-specific feature representation and then fine-tunes it on a small amount of labelled samples. On the other hand, ACPL picks a subset of unlabelled samples based on their informativeness, which improves the data distribution of training subsets. In terms of potential clinical practice, the feature representation from S2MTS2 could be useful for various downstream tasks, and ACPL could result in better classification performance.

Another significant challenge for weakly-supervised learning is the learning with noisy labels. Mainstream methods for noisy label learning employ the small-loss hypothesis, which empirically assumes that training samples with small losses are clean-labelled samples in the early stages of training. However, this hypothesis is incompatible with class imbalance or multi-label scenarios in Medical Image Analysis (MIA) because noisy-label samples from majority classes can have smaller losses than clean-label samples from minority classes. Motivated by these challenges, we developed a robust medical image classification method capable of handling the imbalanced learning problem caused by skewed class distribution and the noisy label problem resulting from unreliable annotations. We proposed a novel method named Non-Volatile Unbiased Memory (NVUM). This method stores a running average of model logits to regularize the training loss and adjusts the class prior distribution to address the imbalanced learning problem. The primary contribution of our method is the incorporation of a non-volatile memory module that can adapt to changing data distribution and label quality, and an unbiased loss function that can mitigate the influence of noisy labels and enhance generalization performance. We evaluated our method using four datasets: ChestX-ray14, CheXpert [75], OpenI [38], and PadChest [16]. Our method outperformed state-of-the-art methods for both imbalanced learning and noisy label learning. We also conducted an ablation study to analyze the impact of each component of our method. The results underscored the effectiveness and robustness of our method for medical image classification under challenging scenarios.

In response to the complex problem of noisy multi-label classification, we proposed an innovative method known as the Bag of Multi-label Descriptors (BoMD). This

method is specifically designed to address the shortcomings of existing techniques that struggle with multi-label scenarios characterized by intricate label dependencies and noise patterns. The BoMD method comprises two primary components: the learning of a bag of multi-label image descriptors (MID) and the smooth re-labelling of images based on a graph structure. The MID captures the nuanced semantic information of each image, proving to be more robust and informative than single global descriptors. The smooth re-labelling scheme employs a graph structure to disseminate label information among similar images, guided by their MID descriptors. We conducted comprehensive experiments on three real-world chest X-ray datasets, each exhibiting different levels and types of label noise. The results demonstrated that the BoMD method surpasses state-of-the-art methods for both noisy and clean labels, thereby validating the effectiveness of each component of our method.

For noisy multi-label learning in medical imaging, NVUM addresses the problem by understanding the negative impact of noisy label gradient and compensates with a specially designed memory module with imbalanced distribution prior. BoMD uses both visual and semantic descriptors for detecting and labeling noisy multi-label images. In terms of clinical practice, NVUM considers both imbalanced learning and noisy label in a unified framework, while BoMD explores multi-modality information and achieves better classification performance.

In traditional multi-class Learning with Noisy Labels (LNL), the relabelling of noisy samples often depends on model predictions as pseudo labels and cross-selection between two independent models. We propose learning from a set of candidate pseudo labels rather than a single , and possibly noisy label. To this end, we introduce a novel method for noisy label learning based on multi-label learning, termed Asymmetric Co-teaching (AsyCo). AsyCo comprises two models: a multi-class classification model n_θ and a multi-label ranking model r_ϕ, which are co-trained using different strategies. The model n_θ selects clean samples based on the prediction disagreement between n_θ and r_ϕ, while r_ϕ relabels noisy samples based on the top-ranked predictions of r_ϕ. We also introduced a multi-view consensus framework that leverages the label views from the training set and the model predictions to estimate the sample selection and relabelling variables. Extensive experiments were conducted on several benchmark datasets with different types of noise, demonstrating that AsyCo outperforms existing state-of-the-art methods in terms of accuracy and robustness. We also provided empirical analysis and ablation studies to validate the effectiveness of our method. Our work underscores the potential of asymmetric co-teaching and multi-view consensus for learning from noisy labels.

We further explored the concept of learning with a noisy label set through partial label learning. We introduced a novel generative method for noisy label learning, termed Generative Noisy-Label Learning by Implicit Discriminative Approximation with Partial Label Prior (GNL). GNL utilizes the implicit discriminative approximation

of the generative model to learn from noisy labels and incorporates a partial label prior to guide the latent label distribution. Additionally, we proposed a coverage and uncertainty measure to monitor the training process and adjust the partial label generation. We conducted experiments on several benchmark datasets with different types of noise, demonstrating that GNL outperforms existing state-of-the-art methods in terms of accuracy and efficiency. We also provided empirical analysis and ablation studies to validate the effectiveness of our method. Our work underscores the potential of generative models and partial label information for learning from noisy labels.

For traditional multi-class noisy label learning in computer vision, both AsyCo and GNL explore the construction of a label set instead of a single label. AsyCo achieves this by training a multi-label ranking model and constructs a label set as a multi-label target. It further proposes a new sample selection based on prediction differences between the multi-label model and the multi-class model. GNL explores a generative learning framework and constructs a label set with a partial label form. For comparison, AsyCo achieves better performance but requires an ensemble of models, while GNL, built with a theoretical guarantee, only requires one model to achieve comparable performance.

9.1 Limitation and Future Work

In this thesis, we have developed novel methods for semi-supervised learning in Medical Image Analysis (MIA). For self-supervised learning in MIA Semi-Supervised Learning (SSL) (Chapter 2.3), the proposed method primarily follows the computer vision strategy without exploring domain-specific strategies for MIA tasks. Moreover, the current experiments do not explore 3D volume images, which are unique tasks for MIA. Future work will involve exploring domain-specific MIA strategies for better self-supervised learning, and better data augmentation. We will also extend our approach to work with 3D volume images to further validate our approach. For Anti-Curriculum Pseudo-Labelling (ACPL) in MIA SSL (Chapter 4), the current selection criterion of informative unlabeled samples does not work if the unlabeled set contains out-of-distribution (OOD) samples. SSL with OOD samples is a practical scenario in real-world datasets, and the future work of ACPL involves developing effective measurements under OOD samples and testing in computer vision benchmarks. A similar idea of ACPL that selects informative samples instead of confidence samples for training semi-supervised learning models has been explored in [191] and we will explore our idea in computer vision domain.

We also concentrated on developing new noisy label methods for real-world MIA tasks and a new perspective from inexact supervision. Currently, Non-Volatile Unbiased Memory (NVUM) (Chapter 5) considers class imbalance distribution when updating the memory module to make the learning robust to imbalanced distribution.

However, pseudo labels also contain class imbalances that should be addressed during training [194]. Moreover, class correlation commonly exists in multi-label classification. In future work of NVUM, we will explore a precise estimation of class prior during the training and class correlation for accurate unbiasing. The limitation of Bag of Multi-label Descriptors (BoMD) (Chapter 6) is the long training time due to multiple stages of training. Another limitation is that class imbalance learning is not explicitly addressed in the current framework. For future work, we will explore integrating the first stage rank loss and relabelling as a joint framework and adapting BoMD to work seamlessly for multi-class and multi-label tasks with minimal adaptation. We would also explore our approach that combines visual descriptor and semantic descriptor for computer vision multi-label noisy label problem, as current approach [208] only explores single image modality.

For Asymmetric Co-teaching (AsyCo) (Chapter 7), the current framework requires two models with different training strategies. We plan to explore lighter models for the multi-label model, as only rank prediction is required. Additionally, Binary Cross-Entropy (BCE) loss for multi-label learning can be solely used in noisy label learning, as noisy samples can benefit from negative labels. Future work will explore better usage of BCE loss to identify relevant label sets for multi-label learning. We will also explore AsyCo under an imbalanced noisy label setup. As AsyCo does not rely on the traditional small-loss assumption for selecting clean samples, this could pose a problem under imbalanced data distribution with majority classes. For Generative Noisy-Label Learning by Implicit Discriminative Approximation with Partial Label Prior (GNL) (Chapter 8), the current approximation is calculated within each batch, and partial label construction still takes a heuristic approach. In future work, we will explore better ways for calculating the approximation for the whole training set and connect partial label construction with conformal predictions, as it is a theoretically guaranteed way for creating label sets with clean label coverage guarantee.

Appendix A

ACPL (Chapter 4) Appendix

A.1 Additional Ablation study

The Number of Information Content Sets in Eq.2 is studied in Table A.1, which shows the model training performance (in terms of mean AUC testing set results) and number of training stages using 2% and 20% labelled set on Chest X-ray14 [192]. The default setting used in the paper is to have three information content sets, namely low, medium, high. As shown in Table A.1, the selection of only two sets produces the worst results because the pseudo-labelled set becomes less informative and imbalanced. The selection of four sets produces similar results as with three sets. However, with this additional set, the number of new pseudo labelled samples are greatly reduced for every training stage, forcing the number of training stages to grow. Hence, by selecting three sets we reach a good balance between training time and accuracy.

A.2 Data Distribution

In Figure A.1, we show the data distribution of all classes of Chest X-ray14 (plus the class 'No Findings') [192]. Notice that the selection of high information content samples (blue) creates a more balanced distribution compared with the selection of low information content (yellow) or the original data distribution (green).

A.3 Visualization of Classification Results

Figure A.2 shows examples of pseudo-labels produced by our density mixup for both Chest Xray-14 [192] (top) and ISIC2018 [182] (bottom) datasets.

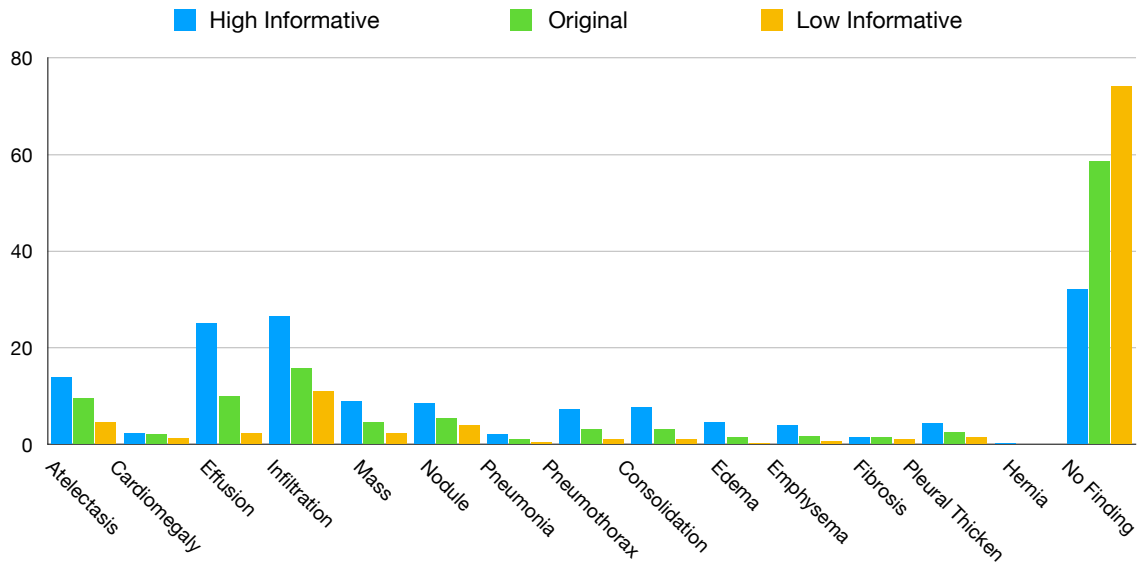


Figure A.1: Histogram of label distribution in percentage of all 14 classes from Chest X-ray14 plus the class 'No Finding'. Blue for high information content subset and yellow for low information content subset. Green is the original data distribution.

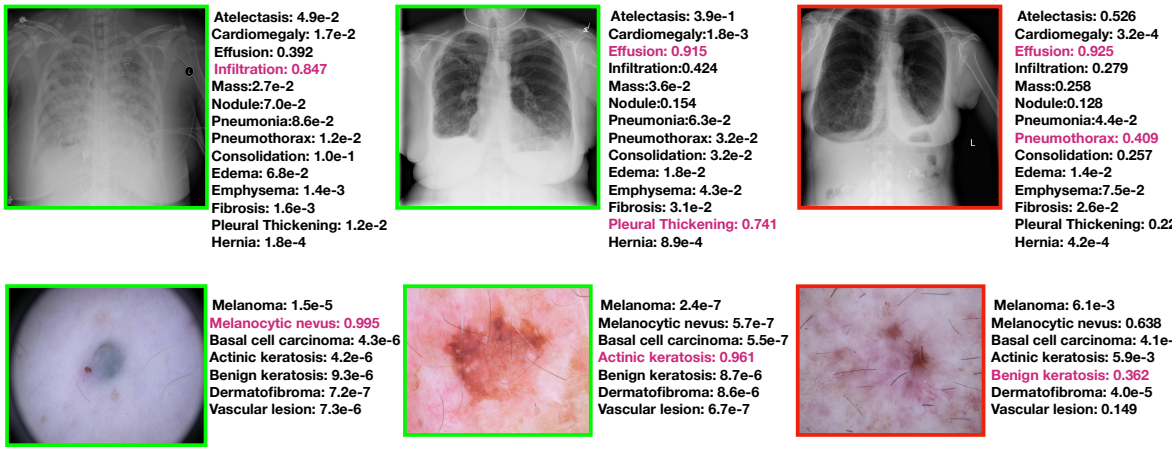


Figure A.2: Pseudo-labelling of high-information content unlabelled samples estimated with the **density mixup** prediction for Chest Xray-14 [192] (top) and ISIC2018 [182] (bottom) datasets. Green border denotes accurate prediction and red border represents inaccurate prediction. Classes with red color represent the ground truth.

Table A.1: Ablation study of the number of information content sets in Eq.2 (2, 3, 4 sets) with model training performance (in terms of mean AUC testing set results) and number of training stages with 2% and 20% labelled set on Chest X-ray14 [192].

Number of Inform. Cont. Sets in Eq.2	2	3	4
Number of Training Stages	5	5	9
2%	71.28	74.44	74.37
20%	79.56	81.51	81.60

Appendix B

NVUM (Chapter 5) Appendix

B.1 Gradient Proof

The gradient for $\ell_{total}(\mathcal{S}, \mathbf{t}, \theta) = \frac{1}{|\mathcal{S}|} \sum_{(\mathbf{x}^i, \tilde{\mathbf{y}}^i) \in \mathcal{S}} \ell_{BCE}(\tilde{\mathbf{y}}^i, \mathbf{p}^i) + \ell_{REG}(\mathbf{t}^i, \mathbf{p}^i)$, is defined as

$$\begin{aligned} \nabla_\theta \ell_{total}(\mathcal{S}, \theta) &= \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbf{J}_{\mathbf{x}^i}(\theta)(\mathbf{p}^i - \tilde{\mathbf{y}}^i + \mathbf{g}^i), \\ \text{where} \quad \mathbf{g}^i_c &= -\sigma((\mathbf{t}^i)^\top(\mathbf{p}^i))\mathbf{p}^i_c(1 - \mathbf{p}^i_c)\mathbf{t}_c, \end{aligned} \tag{B.1}$$

where $\nabla_\theta \ell_{BCE}(\theta) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbf{J}_{\mathbf{x}^i}(\theta)(\mathbf{p}^i - \tilde{\mathbf{y}}^i)$, and recalling that $\ell_{REG}(\mathbf{t}^i, \mathbf{p}^i) = \log(1 - \sigma((\mathbf{t}^i)^\top \mathbf{p}^i))$ (with $\mathbf{p} = \frac{1}{1+e^{-f_\theta(\mathbf{x})}}$), we have

$$\begin{aligned} \nabla_\theta \ell_{REG}(\theta)^i_c &= \mathbf{g}^i_c \\ &= \frac{1}{1 - \sigma((\mathbf{t}^i)^\top(\mathbf{p}^i))} \nabla_\theta (1 - \sigma((\mathbf{t}^i)^\top(\mathbf{p}^i))) \\ &= \frac{-1}{1 - \sigma((\mathbf{t}^i)^\top(\mathbf{p}^i))} \nabla_\theta (\sigma((\mathbf{t}^i)^\top(\mathbf{p}^i))) \\ &= \frac{-1}{1 - \sigma((\mathbf{t}^i)^\top(\mathbf{p}^i))} (\sigma((\mathbf{t}^i)^\top(\mathbf{p}^i)))(1 - \sigma((\mathbf{t}^i)^\top(\mathbf{p}^i))) \nabla_\theta (\mathbf{t}^i)^\top(\mathbf{p}^i) \\ &= -\sigma((\mathbf{t}^i)^\top(\mathbf{p}^i)) \nabla_\theta ((\mathbf{t}^i)^\top(\mathbf{p}^i)) \\ &= -\mathbf{J}_{\mathbf{x}^i}(\theta) \sigma((\mathbf{t}^i)^\top(\mathbf{p}^i))(\mathbf{p}^i_c(1 - \mathbf{p}^i_c)\mathbf{t}^i_c) \end{aligned} \tag{B.2}$$

B.2 Dataset Description

Training Scheme	NIH/CXP	OPI	PDC	Number of Classes
NIH-OPI-PDC	83,672	2,971	14,714	14
CXP-OPI-PDC	170,958	2,823	12,885	8

Table B.1: Statistics of training/testing sets after trimming for consistency between different datasets.

In order to keep the classes consistent between different datasets, we align training set and testing set with the shared classes and exclude unique classes (see Tab. B.1). For further detail, please check [30] for each class.

B.3 Synthetic Label Noise Result

We include results using the public noisy-label medical image benchmark from [234] to test our method for different rates of symmetric noise. We tested our method on their ResNet-18 benchmark, where their baseline accuracy result using the 100% clean set is 64.4%. With 20% symmetric noise, our method without our prior reaches 61.3% and with our prior has 63.1% (best result in [234]: 59.37%). With 40% symmetric noise, our method without our prior reaches 50.7% and with our prior has 53.4% (best result in [234]: 49.65%).

B.4 Memory Footprint

We analysis the memory footprint of our method. In particular, our memory module is a matrix with N x C dimension. We used debugging tools to analyse our memory module and noted that we only required 4 MB of GPU memory, and at each iteration, we only backpropagated through a small subset of the memory.

Appendix C

BoMD (Chapter 6) Appendix

C.1 Dataset Statistics

Table C.1 shows the statistics of our training noisy training set (NIH [193] and ChestX-pert (CXP) [75]) and clean testing sets (OpenI [38] and PadChest [16]). Due to inconsistencies in the number of labels for each dataset, we trim the original datasets and only keep the samples that contain labels present in all datasets based on [31, 114]. After our data pre-processing, there are 83,672 frontal-view images with 14 common chest radiographic observations for NIH [193] dataset where the corresponding testing sets for OpenI [38] and PadChest [16] contain 2,917 and 14,714 frontal-view images respectively. For CXP, we have 170,958 frontal-view images with 8 chest radiographic observations where the corresponding testing set for OpenI [38] and PadChest [16] contain 2,823 and 12,885 frontal-view images, respectively.

C.2 Further Ablation Studies

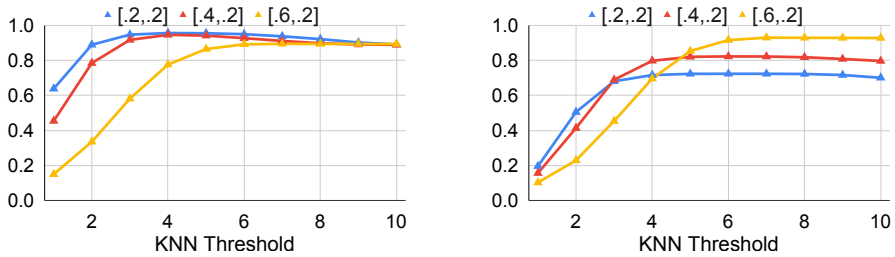
We evaluate the number of KNN neighboring samples that are required for a clean re-labelling. We measure the precision and recall for the detection of noisy-labels of

	Train		Test	
Datasets	NIH [193]	CXP [75]	OpenI [38]	PadChest [16]
Train on NIH	83,672 (14)	-	2,971 (14)	14,714 (14)
Train on CXP	-	170,958 (8)	2,823 (8)	12,885 (8)

Table C.1: Statistics for all datasets after data pre-processing, where the digit on the left is the total number of samples and the digit inside brackets is the number of classes.

Experiments	Mixup Coefficient						Number of Descriptors			K-nearest neighbour		
Settings	λ	OpenI	PadChest	γ	OpenI	PadChest	M	OpenI	PadChest	K	OpenI	PadChest
AUC	0.2	88.39	85.52	0.05	89.14	86.05	1	88.34	86.02	5	89.20	86.15
	0.4	88.56	85.93	0.15	87.87	86.17	3	**89.52**	**86.50**	10	**89.52**	**86.50**
	0.6	**89.52**	**86.50**	0.25	**89.52**	**86.50**	5	88.92	86.39	20	88.23	85.79
	0.8	88.37	86.29	0.35	88.40	86.48	7	89.03	86.43	50	87.59	85.49
	1.0	88.31	86.21	0.45	88.46	86.46	9	88.45	86.29	100	87.36	85.48

Table C.2: Ablation study of the hyper-parameters using mean AUC. Models are trained on NIH [193] and tested on OpenI [38] and PadChest [16]. Note that for each hyper-parameter, we fix the others to their best values (i.e., $\lambda = 0.6$, $\gamma = 0.25$, $M = 3$ and $K = 10$).



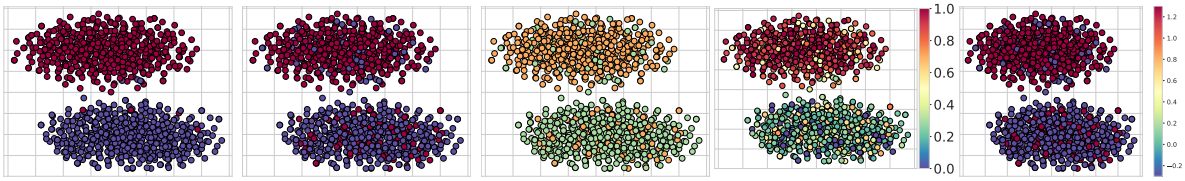
(a) Recall (b) Precision

Figure C.1: Label-wise precision and recall of our KNN propagated label under $\bar{\mathbf{y}}$ w.r.t the clean annotation from PadChest. The horizontal axis shows a threshold of the minimum number of nearest neighbors containing each class.

our graph-based relabelling method in Fig. C.1 as a function of the threshold of the minimum number of nearest neighbors containing each class. For example, if the KNN threshold is 4, then a particular label of a sample is set to 1 only if there are at least 4 neighbors that share the same label. Note that the measures are computed in a label-wise manner, and we consider the flipping rate p_l at 20% and the percentage of noisy samples $p_s \in \{20\%, 40\%, 60\%\}$. We observe a lower recall rate for lower values of K because the KNN label propagation under the multi-label scenario tends to be noisier for small values of K. We achieve the highest recall rate when this threshold is between 4 and 6 nearest neighbours, which means that when we have at least 4 samples in the K nearest neighbour that share the same label, it is most likely a true label.

C.3 Visualisation of Smoothing Techniques

To visualise the performance of different label smoothing techniques, we plot the t-SNE [185] for a toy problem. More specifically, we first generate two isotropic Gaussian clusters as the clean set (Fig. C.2a) and randomly inject 20% of symmetric noise



(a) Clean (b) Noisy (c) LS [128] (d) BoMD (e) GLS [200]

Figure C.2: Visualisation of different label smoothing techniques. The color of each data point indicates the confidence score. We start with two isotropic Gaussian clusters in **(a)** as the clean set where red points indicate class 1 and blue points represent class 2. We randomly inject 20% of symmetric noise to form the noisy set in **(b)**. We compare our method (in **(d)**) with two baseline methods, namely: label smoothing (LS) [128] (in **(c)**) and generalised label smoothing (GLS) [200] (in **(e)**). We show that our method alleviates the noisy label problem by modifying the confidence score based on the nearest neighbors, while LS pushes the labels toward the uniform distribution and GLS pushes the labels toward the sharp binary distribution. Note that GLS has a different scale for confidence scale which is from -0.2 to +1.2, while the others have a range from 0 to 1.

(Fig. C.2b) to form a noisy set. We show that our BoMD demonstrates a better tradeoff when correcting the labels since it re-labels the noisy samples without being overconfident in the detection (like shown by GLS [200]) and without over-smoothing the labels (like displayed by LS [128]). Note that we set the smoothing parameter r to 0.6 and -0.4 respectively for LS [128] and GLS [200].

Table C.3: Disease-level testing AUC results for models **trained on NIH**.

Models	Hermoza et al		CAN		DivideMix		FINE		ELR		NVUM	
Datasets	OpenI	PadChest	OpenI	PadChest	OpenI	PadChest	OpenI	PadChest	OpenI	PadChest	OpenI	PadChest
Atelectasis	86.85	83.59	84.83	79.88	70.98	73.48	77.51	67.70	86.21	85.69	88.16	85.66
Cardiomegaly	89.49	91.25	90.87	91.72	74.74	81.63	77.93	84.54	90.79	92.81	90.57	92.94
Effusion	94.05	96.27	94.37	96.29	84.49	97.75	74.39	86.76	94.74	96.67	93.64	96.56
Infiltration	77.48	70.61	77.88	73.78	84.03	81.61	73.41	67.28	78.92	73.82	74.30	72.51
Mass	95.72	86.93	87.47	85.81	71.31	74.41	57.45	69.54	81.90	84.51	93.06	85.93
Nodule	81.68	75.99	80.71	74.14	57.35	63.89	59.43	57.66	86.22	75.59	88.79	75.56
Pneumonia	87.15	75.73	84.79	76.49	71.65	72.32	56.22	60.46	88.99	80.28	90.90	82.22
Pneumothorax	75.34	74.55	82.21	79.73	75.56	75.46	59.88	64.46	78.65	78.47	85.78	79.50
Edema	84.31	97.78	82.80	96.41	80.71	85.81	58.18	95.20	85.57	97.58	86.56	95.70
Emphysema	83.26	79.81	81.26	78.06	64.81	59.91	43.31	50.72	82.79	79.87	83.70	79.38
Fibrosis	85.85	96.46	83.17	93.20	76.96	84.71	61.97	88.68	92.07	97.42	91.67	97.61
Pleural Thicken	77.99	71.85	77.59	67.87	62.98	58.25	63.17	54.33	83.45	72.01	84.82	74.80
Hernia	92.90	89.90	87.37	86.87	70.34	72.11	64.86	74.56	95.77	93.37	94.28	93.02
Mean AUC	85.54	83.90	84.26	83.10	72.76	75.49	63.67	70.91	86.62	85.24	88.17	85.49

Table C.4: Disease-level testing AUC results for models **trained on NIH**.

Models	NPC		NCR		LS		OLS		GLS		**BoMD**	
Datasets	OpenI	PadChest	OpenI	PadChest	OpenI	PadChest	OpenI	PadChest	OpenI	PadChest	OpenI	PadChest
Atelectasis	86.04	85.23	83.80	85.46	85.34	84.74	87.27	85.18	88.23	83.00	87.91	86.19
Cardiomegaly	91.42	92.12	89.42	91.45	88.08	89.17	84.59	89.83	89.12	91.40	91.37	92.17
Effusion	95.58	96.19	93.96	95.89	94.54	95.63	94.28	96.75	93.67	96.36	95.28	96.71
Infiltration	68.76	64.08	60.48	67.98	72.26	74.20	76.10	76.19	82.08	71.27	81.65	76.64
Mass	80.20	86.04	85.00	85.98	88.08	80.56	82.79	84.80	75.12	80.67	92.31	88.48
Nodule	87.60	75.68	85.12	75.60	86.44	74.82	83.42	75.27	82.10	74.34	84.05	75.28
Pneumonia	91.01	76.87	88.87	76.40	83.50	76.17	87.18	78.20	85.65	74.83	89.99	78.71
Pneumothorax	84.28	79.22	83.07	76.98	74.07	76.10	75.89	80.02	73.93	76.45	88.89	85.82
Edema	82.27	92.40	85.66	93.87	83.38	88.23	87.31	89.55	85.92	93.01	87.60	98.68
Emphysema	82.05	80.87	82.36	75.80	76.94	73.10	80.94	78.15	75.16	74.21	85.28	81.94
Fibrosis	87.53	91.50	90.67	94.57	92.09	96.43	90.19	95.35	91.06	95.29	94.56	97.44
Pleural Thicken	87.37	76.06	82.66	76.62	82.83	72.82	84.12	70.55	80.10	68.14	86.94	71.53
Hernia	96.60	94.17	94.69	92.74	80.85	70.11	91.95	85.84	87.29	81.38	98.57	94.22
Mean AUC	86.21	83.88	85.06	83.79	83.72	80.93	85.08	83.51	83.80	81.56	89.57	86.45

C.4 Additional Results

C.4.1 Per-finding results

We show per-finding results over all available findings for NIH [193] in Tables C.3 and C.4 and for CheXpert [75] in Tables C.5 and C.6 .

C.4.2 Hyper-parameter sensitivity

Table C.2 studies the four hyper-parameters (λ, γ, M and K) of BoMD. In general, for λ, we note that relying too much on the pseudo-labels from the graph ($\lambda = 0.2$) or the original noisy labels ($\lambda = 1.0$) worsens the performance, with the best result achieved with a balanced $\lambda = 0.6$. We noticed that the method is robust to γ and M with little variation in results. As for K, values larger than 10 over-smooth the decision boundary of our classifier, causing under-fitting. The values $\lambda = 0.6$ and $\gamma = 0.25$, $M = 3$, and $K = 10$ reach the best results.

C.4.3 Evaluation for Descriptors from MID

Visualisation of distance distribution. To verify the separation of positive descriptors (labelled as 1) and negative descriptors (labelled as 0) based on their edge weight, we performed an analysis on a dataset consisting of 12 classes. Each class contained 4,000 samples, along with its corresponding semantic descriptors from the NIH dataset [193]. For each class, we denote positive samples' descriptors as "1", and negative samples' descriptors as "0". The analysis involved examining the distribution of L2 distance, and the results are presented in Figure C.3. Our findings suggest that, on average, positive descriptors are closer to their corresponding semantic descriptors than negative descriptors, which proves the effectiveness of our MID module.

Visualisation of latent space. To visualise the descriptors' distribution in the latent space, we plot the t-SNE [185] for 12 classes with 4,000 samples per class sampled from NIH [193], as shown in Fig. C.4. For each class, we denote positive samples' descriptors as $+$, negative samples' descriptors as $\circ$ and semantic descriptors as $\times$. We show that the semantic descriptors are mostly surrounded by class-related descriptors ($+$), which varied the clustering effect of our MID module. Such clustering effect will benefit our graph-based smooth re-labelling as shown in Sec C.3

Table C.5: Disease-level testing AUC results for models that **trained on CheXpert**.

Models	Hermoza et al		CAN		DivideMix		FINE		ELR		NVUM	
Datasets	OpenI	PadChest	OpenI	PadChest	OpenI	PadChest	OpenI	PadChest	OpenI	PadChest	OpenI	PadChest
Cardiomegaly	86.12	87.20	82.83	85.89	79.53	85.42	83.62	83.99	90.48	87.46	85.15	88.48
Edema	87.92	94.35	86.46	97.47	81.24	83.41	86.43	87.07	90.88	96.12	87.35	97.21
Pneumonia	65.56	57.15	61.88	63.38	55.98	51.20	55.58	55.58	61.59	64.13	64.42	67.89
Atelectasis	78.40	75.65	80.13	72.87	72.74	68.34	72.87	72.87	79.63	73.68	80.81	75.03
Pneumothorax	62.09	78.65	74.69	79.50	75.49	79.98	65.34	68.85	74.12	83.95	82.18	83.32
Effusion	87.00	93.94	88.43	92.92	83.75	88.91	85.92	85.92	86.65	92.42	83.54	89.74
Fracture	57.47	53.77	59.96	60.44	63.87	62.23	51.97	62.50	56.75	62.00	57.02	62.67
Mean AUC	74.94	77.24	76.34	78.92	73.23	74.21	71.68	73.83	77.16	79.97	77.21	80.62

Table C.6: Disease-level testing AUC results for models that **trained on CheXpert**.

Models	NPC		NCR		LS		OLS		GLS		**BoMD**	
Datasets	OpenI	PadChest	OpenI	PadChest	OpenI	PadChest	OpenI	PadChest	OpenI	PadChest	OpenI	PadChest
Cardiomegaly	80.33	86.43	90.10	86.84	85.53	83.42	83.58	86.29	88.22	87.30	90.85	89.88
Edema	82.35	79.09	90.11	98.26	89.72	99.43	85.17	95.69	87.92	97.49	89.89	98.76
Pneumonia	62.31	64.52	58.80	59.87	49.64	50.41	64.18	56.48	59.49	63.64	65.35	66.10
Atelectasis	81.29	76.13	79.01	72.22	75.13	69.30	70.85	71.75	76.71	73.32	80.01	74.33
Pneumothorax	82.32	82.35	78.06	86.15	73.05	78.33	80.10	83.36	77.53	77.58	82.99	86.04
Effusion	78.71	86.65	85.62	91.57	84.70	90.97	84.64	91.83	85.19	91.94	87.37	93.07
Fracture	59.92	65.95	56.80	60.63	52.27	55.52	67.13	58.60	60.44	60.32	63.72	64.12
Mean AUC	75.32	77.30	76.93	79.36	72.86	75.34	76.52	77.72	76.50	78.80	80.03	81.76

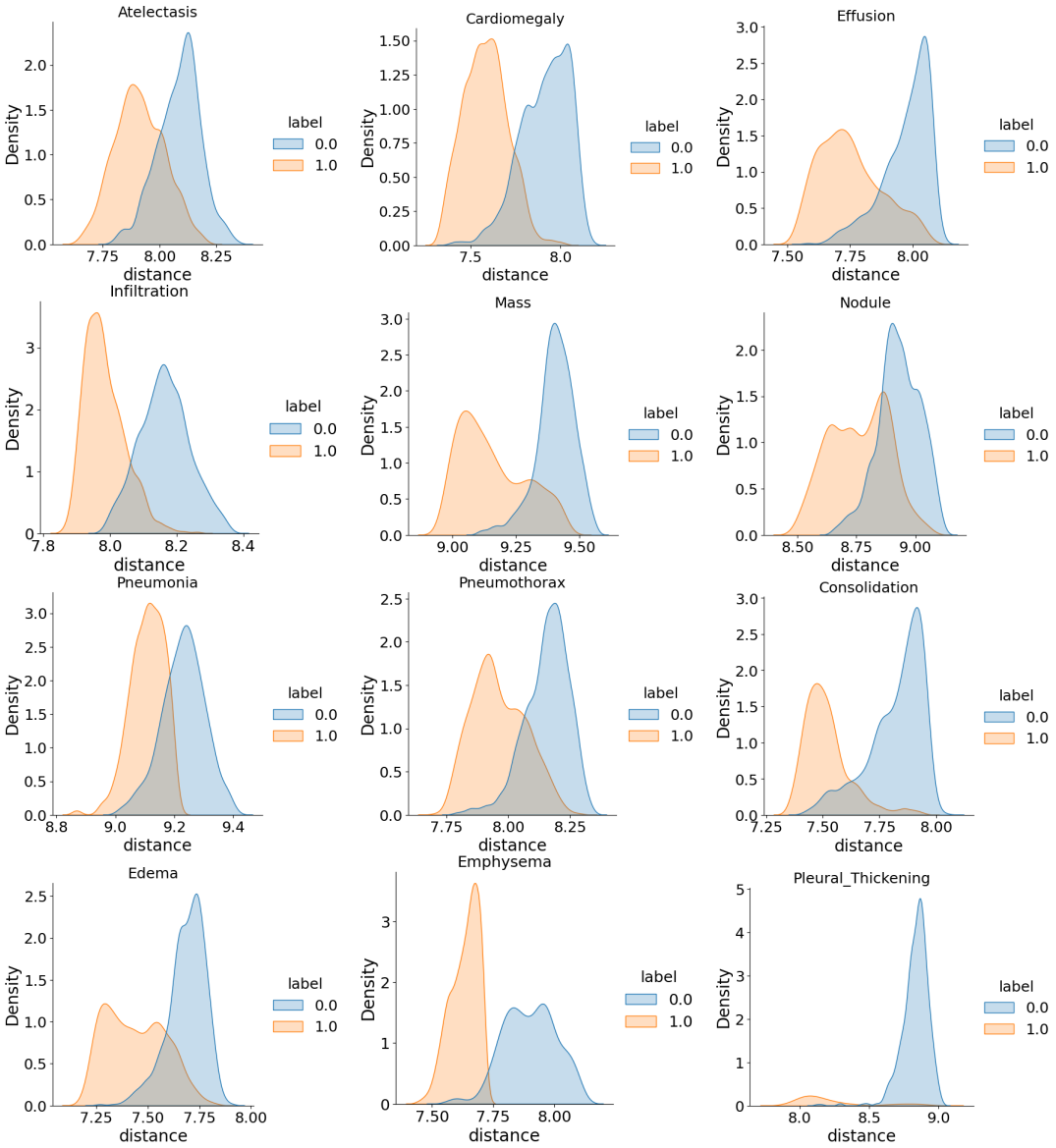


Figure C.3: L2 distance between positive/negative descriptors and semantic descriptor

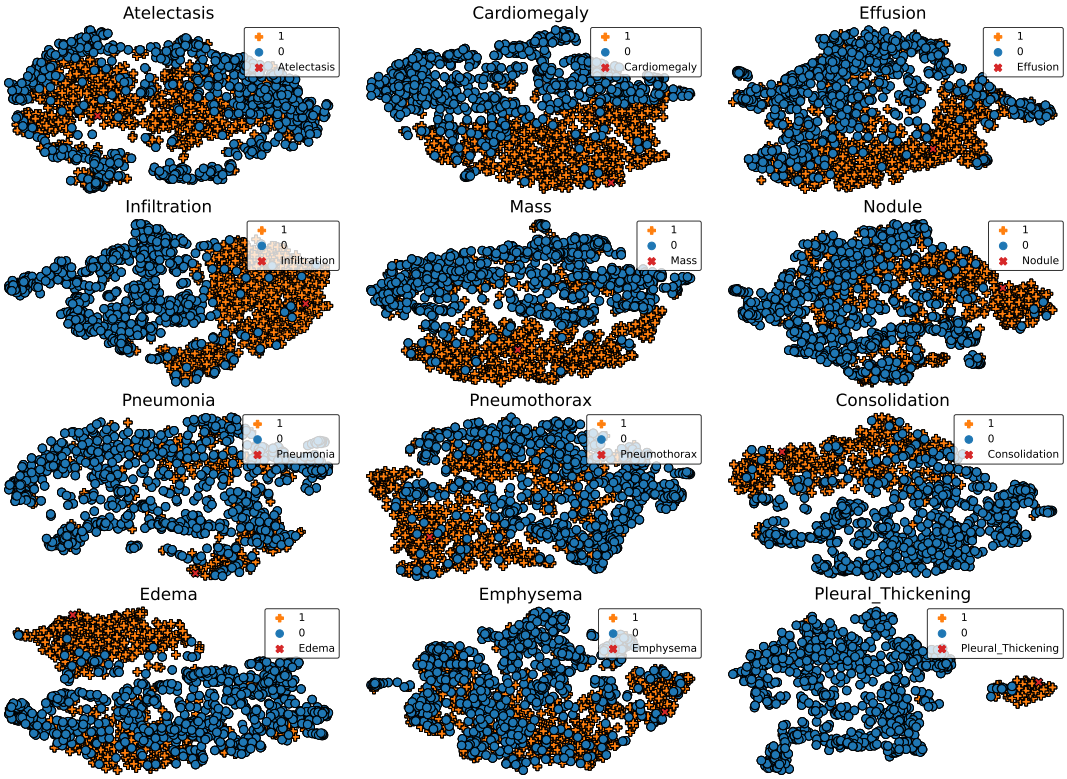
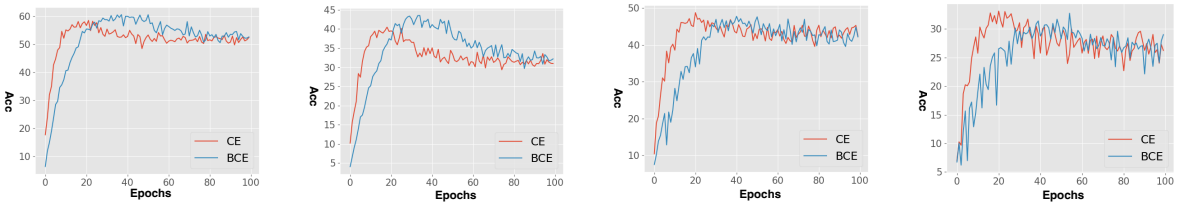


Figure C.4: Visualisation of descriptor distribution in latent space.

Appendix D

AsyCo (Chapter 7) Appendix



(a) CIFAR100 0.2 Acc (b) CIFAR100 0.5 Acc (c) Red Mini 0.2 Acc (d) Red Mini 0.8 Acc

Figure D.1: Comparison of the accuracy between a model trained with CE loss and another trained with BCE loss. The comparison is done for a training that lasts 100 epochs on CIFAR100 with 0.2/0.5 noise rates and Red Mini-ImageNet with 0.2/0.8 noise rates.

D.1 AsyCo Training Algorithm

Algorithm 3 shows the main steps of our proposed AsyCo training algorithm.

D.2 Training Strategy Visualization

In Fig. D.1, we show a visualisation of the testing accuracy differences for training two models, one with with CE loss and another with BCE loss, for 100 epochs on instance-dependent CIFAR100 with 0.2/0.5 noise rates and real-world Red Mini-ImageNet with 0.2/0.8 noise rates. We observe that CE and BCE show distinct training behaviours on both datasets and noise rates. Training with CE converges faster than BCE, but it also overfits more easily than BCE. Training with BCE takes longer to reach the same performance as CE, but it also overfits more slowly. This suggests that differences in training strategies can be explored for multi-view consensus selection.

Algorithm 3 Asymmetric Joint Training Algorithm

require: Training net $n_\theta(.)$, reference net $r_\phi(.)$, training dataset $\mathcal{D}$ and number of training epochs T
Warm-up $n_\theta(.)$ and $r_\phi(.)$ with Eq. 1
while $t < T$ **do**
  Compute $\tilde{\mathbf{y}}_i^{(n)}$ and $\tilde{\mathbf{y}}_i^{(r)}$ with Eq. 2
  Categorize training samples from Tab. 1
  Estimate sample selection latent variable $\mathbf{w}$, from Eq. 4, which classifies samples into clean or noisy
  Train $n_{\theta^*}(.)$ with Eq. 5 using $\mathbf{w}$ and $\mathcal{D}$
  Estimate latent variable $\hat{\mathbf{y}}$, from Eq. 6, which re-labels the training samples with multiple labels
  Train $r_{\phi^*}(.)$ with Eq. 7 using $\hat{\mathbf{y}}$ and $\mathcal{D}$
end while
return n_{θ^*}

Methods	GMM [106]	FINE [91]	Ours
Time	17.2s	34s	13s

Table D.1: Time differences for each sample selection strategy on CIFAR10.

D.3 Sample Selection Time Comparison

In Tab. D.1, we show the running times of the small-loss based selection (GMM from DivideMix [106]), eigenvector-based selection (FINE [91]), and our multi-view consensus selection. We observe that our approach is the most efficient, being 4s faster than small-loss and 20s faster than FINE. The reason is that our approach utilizes bit-wise AND operation for three views to partition the large training set into multiple subsets, which avoids multiple EM estimations. FINE performs the slowest because of the complexity involved in estimating the eigenvectors for each class.

D.4 Ablation Study of Hyper-parameters K and λ

Table D.2 shows accuracy (%) on CIFAR100 instance-dependent noise as a function of K and λ. We observe that the model accuracy is relatively robust to a wide range of values of K and λ, but in general, accuracy drops when K increases and is stable as a function of λ.

CIFAR100 (Inst. Depend. Noise)	0.2	0.3	0.4	0.5
Paper result (K =3,λ=100)	76.02	74.02	68.96	60.35
K =5	75.8	73.9	68.3	58.8
K =10	74.7	72.5	67.5	57.7
K =20	74.3	71.5	65.8	56.4
λ=25	75.8	73.4	67.4	59.3
λ=50	75.6	73.3	68.5	60.2
λ=150	76.0	73.9	68.7	60.1

Table D.2: Test accuracy (%) on CIFAR100 instance-dependent noise hyper-parameter sensitivity test. K is the top-K prediction and λ is unsupervised loss weight.

Appendix E

GNL (Chapter 8) Appendix

E.1 Implementation details

In Tab. E.1, we describe the implementation details of our method for each dataset, including the main reference for the respective baseline. In addition, for the Clothing1M, we sample 1000 mini-batches from the training set, where in every mini batch we ensure that the 14 classes are evenly sampled to form a pseudo-balanced learning problem. Also for Clothing1M, we first resize the image to 256 $\times$ 256 and then random crop to 224 $\times$ 224 with random horizontal flipping. The number of epochs for warmup is 10 for datasets containing 10 classes, and 15 for datasets containing 100 classes. For Clothing1M, we run 1 epoch of warmup.

	IDN CIFAR10/100	CIFAR10/100 N	Red Mini-ImageNet	Animal-10N	Clothing1M
Baseline reference	kMEIDTM [26]	Real-world [203]	FaMUS [216]	Nested [25]	CausalNL [220]
Backbone	ResNet-34	ResNet-34	Pre-act ResNet-18	VGG-19BN	ResNet-50 *
# Training epochs	150	120	150	100	40
Batch size	128	128	128	128	64
Learning rate	0.02	0.02	0.02	0.02	0.002
Weight decay	5e-4	5e-4	5e-4	5e-4	1e-3
LR decay at epochs	0.1/100	0.1/80	0.1/100	0.1/50	0.1/20
Data augmentation	Random Crop / Random horizontal flip				
β	0.9				
K	1				

Table E.1: Implementation detail of our method in each dataset. *: Uses ImageNet pre-trained model.

E.2 Additional ablation study

In Tab. E.2, we perform a hyper-parameter sensitivity test for our method on CIFAR10-IDN, including coverage and uncertainty for prior label construction. To test label

		CIFAR10			
		20%	30%	40%	50%
$\beta = 0.9$, $K = 1$		92.65	91.96	91.02	89.94
Hyper-parameter	$\beta = 0.8$	92.49	91.88	90.83	88.81
	$\beta = 0.7$	91.55	90.87	90.62	88.40
	$\beta = 0.5$	89.13	87.98	87.48	85.73
	$K = 3$	92.30	91.83	90.83	89.75
Coverage	$\beta = 0$	84.57	81.59	68.88	61.47
	arg max	20.19	18.56	16.09	15.26
	No Cov	85.57	81.00	72.42	66.61
Uncertainty	Uniform w	90.10	89.66	86.25	84.28
	No Unc	84.96	83.19	81.88	78.38

Table E.2: Ablation study on hyper-parameter sensitivity, including β, K, coverage and uncertainty.

coverage, we first examine the model performance as function of $\beta \in \{0.5, 0.7, 0.8, 0.9\}$ in Eq. 11, where the default value for $\beta = 0.9$. We observe that performance does not change much for $\beta \in \{0.7, 0.8, 0.9\}$, which indicates our model's robustness with respect to that hyper-parameter. For $\beta = 0.5$, the performance drops significantly, indicating that using a moving average with a relatively high value for β is important for estimating model prediction and avoiding overfitting. We test $K \in \{1, 3\}$ in Eq. 13 by sampling multiple times $\{\hat{\mathbf{y}}_{i,j}\}_{j=1}^{K} \sim \mathsf{Cat}(g_\theta(\mathbf{x}_i))$. We observe no significant changes to model performance with this higher value of K. Therefore, we choose $K = 1$ for simplicity.

We also test our model by shutting down the moving average, either by making $\beta = 0$ or by completely relying on the model's current prediction, which is an experiment denoted by "arg max" in Tab. E.2. Note that the model shows a major performance drop in both cases. This happens because the model overfits to the inaccurate model predictions, biasing the training procedure. Furthermore, arg max performs the worst because inaccurate model prediction in prior label in the early stage, causing confirmation bias and leading to wrong optimisation goal. We also test our model without using the coverage term $\mathbf{c}_i$ in Eq. 10 – this experiment is denoted as "No Cov". In this case, the performance of the model drops significantly for all noise rates, compared to the default model in the first row, which indicates the importance of having a coverage term in our prior label construction.

Furthermore, we study the uncertainty aspect for the prior label construction. We first experiment by setting w_i to a uniform value ("Uniform w" row) instead of a GMM weight that represents the probability that the sample is carrying a clean label. The result is not competitive with the one that uses the GMM weight, which indicates the

importance of having w representing a clean-label sample probability. We also test our model without the uncertainty component $\mathbf{u}_i$ in Eq. 10 (see row "No Unc"). This case shows a significant performance drop in all noise rates.

Bibliography

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In International conference on machine learning, pages 312–321. PMLR, 2019. 3, 13, 87, 88, 91, 104, 106

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In 2020 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE, 2020. 33

[3] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 233–242. JMLR. org, 2017. 104, 106

[4] Angelica I Aviles-Rivero, Nicolas Papadakis, Ruoteng Li, Philip Sellars, Qingnan Fan, Robby T Tan, and Carola-Bibiane Schönlieb. Graphxˆ\small net-net-chest x-ray classification under extreme minimal supervision. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22, pages 504–512. Springer, 2019. 13, 20, 24, 25, 26, 27, 32, 35, 41, 42, 43

[5] HeeSun Bae, Seungjae Shin, Byeonghu Na, JoonHo Jang, Kyungwoo Song, and Il-Chul Moon. From noisy prediction to true label: Noisy prediction calibration via generative model, 2022. xii, xiv, 14, 64, 65, 67, 74, 75, 76, 104, 105, 106, 107, 109

[6] Dara Bahri, Heinrich Jiang, and Maya Gupta. Deep k-nn for noisy labels. In International Conference on Machine Learning, pages 540–550. PMLR, 2020. 67

[7] Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac mr image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 253–260. Springer, 2017. 35, 43, 44

[8] Avi Ben-Cohen, Nadav Zamir, Emanuel Ben-Baruch, Itamar Friedman, and Lihi

Zelnik-Manor. Semantic diversity learning for zero-shot multi-label classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 640–650, 2021. 65

[9] Yoshua Bengio, Jerome Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Proceedings of the 26th annual international conference on machine learning, pages 41–48, 2009. 34

[10] Kristin Bennett and Ayhan Demiriz. Semi-supervised support vector machines. Advances in Neural Information processing systems, 11, 1998. 2

[11] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In International Conference on Learning Representations, 2019. 2, 5, 12, 20, 32, 34

[12] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. Advances in neural information processing systems, 32, 2019. 2, 5, 12, 32, 34, 67, 94

[13] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In Proceedings of the eleventh annual conference on Computational learning theory, pages 92–100, 1998. 87, 88, 89

[14] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In Proceedings of the 22nd international conference on Machine learning, pages 89–96, 2005. 71

[15] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis, 66:101797, 2020. viii, 16, 53, 55, 57, 58

[16] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis, 66:101797, 2020. x, 66, 74, 75, 120, 131, 132

[17] Qi Cai, Yu Wang, Yingwei Pan, Ting Yao, and Tao Mei. Joint contrastive learning with infinite possibilities. arXiv preprint arXiv:2009.14776, 2020. 20, 22, 23, 27

[18] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In Advances in Neural Information Processing Systems, pages 1567–1578, 2019. 53

[19] Paola Cascante-Bonilla et al. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. arXiv preprint arXiv:2001.06001, 2020. 34

[20] Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In International Conference on Machine Learning, pages 1062–1070. PMLR, 2019. 3, 106

[21] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International

conference on machine learning, pages 1597–1607. PMLR, 2020. 5, 13

[22] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020. 20, 22, 25

[23] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems, 33:22243–22255, 2020. 13, 20, 22, 25

[24] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020. 25, 26, 27

[25] Yingyi Chen and et al. Boosting co-teaching with compression regularization for label noise. In CVPR, pages 2688–2692, 2021. ix, 16, 98, 99, 100, 112, 115, 143

[26] De Cheng and et al. Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16630–16639, 2022. ix, 3, 97, 98, 99, 104, 106, 113, 114, 116, 143

[27] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In International Conference on Learning Representations, 2021. 8, 114

[28] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical image analysis, 54:280–296, 2019. 22

[29] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368, 2019. 16, 20, 24, 28, 34, 35, 41

[30] Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRayVision: A library of chest X-ray datasets and models. https://github.com/mlmed/torchxrayvision, 2020. 56, 130

[31] Joseph Paul Cohen, Joseph D Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, et al. Torchxrayvision: A library of chest x-ray datasets and models. arXiv preprint arXiv:2111.00595, 2021. 131

[32] Guillaume Couairon, Matthijs Douze, Matthieu Cord, and Holger Schwenk. Embedding arithmetic of multimodal queries for image retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4950–4958, 2022. 65

[33] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. Workshop on statistical learning in computer vision, ECCV, 1(1-22):1–2, 2004. 68

[34] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pages 702–703, 2020. 12

[35] Wenhui Cui, Yanlin Liu, Yuxing Li, Menghao Guo, Yiming Li, Xiuli Li, Tianle Wang, Xiangzhu Zeng, and Chuyang Ye. Semi-supervised brain lesion segmentation with an adapted mean teacher model. In International Conference on Information Processing in Medical Imaging, pages 554–565. Springer, 2019. 22, 35

[36] Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Russ R Salakhutdinov. Good semi-supervised learning that requires a bad gan. Advances in neural information processing systems, 30, 2017. 8

[37] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association, 23(2):304–310, 2016. viii, 16, 53, 55, 57, 58

[38] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association, 23(2):304–310, 2016. x, 74, 75, 81, 120, 131, 132

[39] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22, 1977. 38, 111

[40] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009. 2, 5, 86, 93, 94, 104, 119

[41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018. 7, 86, 104

[42] Andres Diaz-Pinto, Adrián Colomer, Valery Naranjo, Sandra Morales, Yanwu Xu, and Alejandro F Frangi. Retinal image synthesis and semi-supervised learning for glaucoma assessment. IEEE transactions on medical imaging, 38(9):2211–2218, 2019. 43, 44

[43] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE international

conference on computer vision, pages 1422–1430, 2015. 13
[44] Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. Advances in Neural Information Processing Systems, 34:30284–30297, 2021. 14, 67
[45] Rajpurkar et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. PLoS, 2018. 2
[46] Nicolas Fiorini, Robert Leaman, David J Lipman, and Zhiyong Lu. How user intelligence is improving pubmed. Nature biotechnology, 36(10):937–945, 2018. xiii, 69, 70
[47] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In International conference on machine learning, pages 1183–1192. PMLR, 2017. 2
[48] Arpit Garg and et al. Instance-dependent noisy label learning via graphical modelling. WACV, 2022. 14, 88
[49] Arpit Garg, Cuong Nguyen, Rafael Felix, Thanh-Toan Do, and Gustavo Carneiro. Instance-dependent noisy label learning via graphical modelling. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 2288–2298, 2023. xiv, 104, 105, 106, 107, 109, 115, 116
[50] Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In Thirty-First AAAI Conference on Artificial Intelligence, 2017. 3
[51] Aritra Ghosh, Naresh Manwani, and PS Sastry. Making risk minimization tolerant to label noise. Neurocomputing, 160:93–107, 2015. 6, 8
[52] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728, 2018. 13
[53] Ross Girshick. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 1440–1448, 2015. 1
[54] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In International conference on learning representations, 2016. 14, 52, 65, 67
[55] Chen Gong, Tongliang Liu, Yuanyan Tang, Jian Yang, Jie Yang, and Dacheng Tao. A regularization approach for instance-based superset label learning. IEEE transactions on cybernetics, 48(3):967–978, 2017. 4
[56] Xiuwen Gong, Jiahui Yang, Dong Yuan, and Wei Bao. Generalized large margin k nn for partial label learning. IEEE Transactions on Multimedia, 24:1055–1066, 2021. 4
[57] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. Advances in neural information processing systems, 17, 2004. 12
[58] Qingji Guan and Yaping Huang. Multi-label chest x-ray image classification via category-wise residual attention learning. Pattern Recognition Letters, 130:259–266, 2020. 27

[59] Prashnna Kumar Gyawali, Sandesh Ghimire, Pradeep Bajracharya, Zhiyuan Li, and Linwei Wang. Semi-supervised medical image classification with global latent mixing. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 604–613. Springer, 2020. 20, 24, 26, 28

[60] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In Advances in neural information processing systems, pages 8527–8537, 2018. 3, 8, 13, 52, 56, 58, 67, 76, 87, 88, 91, 97, 98, 99, 100, 104, 106, 113, 114

[61] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. arXiv preprint arXiv:2102.02051, 2021. 89

[62] Zellig S Harris. Distributional structure. Word, 10(2-3):146–162, 1954. 68

[63] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. arXiv preprint arXiv:1911.05722, 2019. 54, 59

[64] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9729–9738, 2020. vii, 20, 22, 23, 24, 28

[65] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9729–9738, 2020. 5, 13

[66] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017. 1

[67] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learningfor image recognition. ComputerScience, 2015. 1, 86, 94, 104

[68] Renato Hermoza et al. Region proposals for saliency map refinement for weakly-supervised disease localisation and classification. arXiv preprint arXiv:2005.10550, 2020. 26, 27, 56, 57, 58, 65, 66, 74, 75, 76, 80

[69] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015. 12

[70] Hexiang Hu, Wei-Lun Chao, and Fei Sha. Learning answer embeddings for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5428–5436, 2018. 65

[71] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7132–7141, 2018. 1

[72] Gao Huang et al. Densely connected convolutional networks. In CVPR, pages 4700–4708, 2017. 25, 41, 43, 56, 57, 74

[73] Jinchi Huang et al. O2u-net: A simple noisy label detection approach for deep neural networks. In ICCV, pages 3326–3334, 2019. 67

[74] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342, 2019. 65, 79, 80

[75] Jeremy Irvin et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In AAAI, volume 33, pages 590–597, 2019. viii, 65, 66, 72, 74, 75, 76, 79, 119, 120, 131, 134

[76] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 590–597, 2019. viii, 15, 16, 20, 24, 27, 35, 52, 53, 55, 58

[77] Ahmet Iscen, Jack Valmadre, Anurag Arnab, and Cordelia Schmid. Learning with neighbor consistency for noisy labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4672–4681, 2022. xii, 64, 65, 66, 68, 75, 76

[78] Angela H Jiang, Daniel L-K Wong, Giulio Zhou, David G Andersen, Jeffrey Dean, Gregory R Ganger, Gauri Joshi, Michael Kaminksy, Michael Kozuch, Zachary C Lipton, et al. Accelerating deep learning by focusing on the biggest losers. arXiv preprint arXiv:1910.00762, 2019. 34

[79] Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. In International conference on machine learning, pages 4804–4815. PMLR, 2020. 15, 93, 94, 99, 115

[80] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. Self-paced curriculum learning. In Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015. 34

[81] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In International Conference on Machine Learning, pages 2304–2313. PMLR, 2018. 3, 15, 106, 112

[82] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In International conference on machine learning, pages 2304–2313. PMLR, 2018. 58, 76, 88, 97, 98

[83] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. Scientific data, 3(1):1–9, 2016. xiii, 69, 70

[84] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3):535–547, 2019. 74

[85] Jeff Johnson et al. Billion-scale similarity search with gpus. arXiv preprint arXiv:1702.08734, 2017. 42

[86] Yaqub Jonmohamadi et al. Automatic segmentation of multiple structures in knee arthroscopy using deep learning. IEEE Access, 8:51853–51861, 2020. 20

[87] Timo Kaiser, Lukas Ehmann, Christoph Reinders, and Bodo Rosenhahn. Blind knowledge distillation for robust image classification. arXiv preprint arXiv:2211.11355, 2022. 14, 106

[88] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. arXiv preprint arXiv:1910.09217, 2019. 53

[89] Kenji Kawaguchi and Haihao Lu. Ordered sgd: A new stochastic optimization framework for empirical risk minimization. In International Conference on Artificial Intelligence and Statistics, pages 669–679. PMLR, 2020. 34

[90] Jang-Hyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song. Co-: Saliency guided joint mixup with supermodular diversity. arXiv preprint arXiv:2102.03065, 2021. 12, 46

[91] Taehyeon Kim, Jongwoo Ko, JinHwan Choi, Se-Young Yun, et al. Fine samples for learning with noisy labels. Advances in Neural Information Processing Systems, 34:24137–24149, 2021. 14, 65, 67, 75, 76, 106, 140

[92] Diederik Pieter Kingma. Variational inference & deep learning: A new synthesis. 2017. 111

[93] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 41, 42, 56, 74

[94] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. Advances in neural information processing systems, 32, 2019. 2

[95] Jan M Köhler, Maximilian Autenrieth, and William H Beluch. Uncertainty based detection and relabeling of noisy image labels. In CVPR Workshops, pages 33–37, 2019. 14, 106

[96] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5, 15, 93, 112

[97] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6):84–90, 2017. 86, 104

[98] M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In Advances in Neural Information Processing Systems, pages 1189–1197, 2010. 34

[99] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242, 2016. 11, 20, 22, 34, 43, 44

[100] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4):541–551, 1989. 1

[101] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998. 1

[102] Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In Proceedings of 2010 IEEE international symposium on circuits and systems, pages 253–256. IEEE, 2010. 1

[103] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. 2013. 2, 5, 12, 20, 32, 34

[104] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240, 2020. 7, 65, 79, 80

[105] June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. Deep learning in medical imaging: general overview. Korean journal of radiology, 18(4):570, 2017. 20

[106] Junnan Li and et al. Dividemix: Learning with noisy labels as semi-supervised learning. ICLR, 2020. xii, xiii, 3, 13, 52, 56, 57, 58, 64, 65, 67, 74, 75, 76, 77, 87, 88, 91, 94, 96, 97, 98, 99, 100, 104, 106, 107, 110, 112, 114, 115, 116, 140

[107] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In International Conference on Learning Representations, 2019. 3, 7, 15

[108] Shikun Li, Xiaobo Xia, Hansong Zhang, Yibing Zhan, Shiming Ge, and Tongliang Liu. Estimating noise transition matrix with label correlations for noisy multi-label learning. Advances in Neural Information Processing Systems, 35:24184–24198, 2022. 65, 68

[109] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. arXiv preprint arXiv:1708.02862, 2017. 2, 3

[110] Xuefeng Li, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. Provably end-to-end label-noise learning without anchor points. arXiv preprint arXiv:2102.02400, 2021. 14, 114

[111] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, and Pheng-Ann Heng. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. arXiv preprint arXiv:1808.03887, 2018. 35, 43, 44

[112] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8290–8299, 2018. 20, 27

[113] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image

analysis. Medical image analysis, 42:60–88, 2017. 32, 52, 63, 86, 104
[114] Fengbei Liu, Yuanhong Chen, Yu Tian, Yuyuan Liu, Chong Wang, Vasileios Belagiannis, and Gustavo Carneiro. Nvum: Non-volatile unbiased memory for robust medical image classification. arXiv e-prints, pages arXiv–2103, 2021. 65, 66, 67, 68, 74, 75, 76, 79, 80, 81, 131
[115] Fengbei Liu et al. Self-supervised depth estimation to regularise semantic segmentation in knee arthroscopy. In MICCAI, pages 594–603. Springer, 2020. 20
[116] Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, and Gustavo Carneiro. Acpl: Anti-curriculum pseudo-labelling forsemi-supervised medical image classification. arXiv preprint arXiv:2111.12918, 2021. 52, 74
[117] Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. In International Conference on Learning Representations, 2021. 5, 32, 34, 35, 41, 42, 43
[118] Quande Liu, Lequan Yu, Luyang Luo, Qi Dou, and Pheng Ann Heng. Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE transactions on medical imaging, 39(11):3429–3440, 2020. 13, 16, 20, 22, 24, 25, 26, 27, 28, 35, 41, 42, 43, 44
[119] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. Advances in neural information processing systems, 33:20331–20342, 2020. xii, 3, 6, 13, 14, 52, 58, 64, 65, 67, 75, 76, 87, 88, 99, 110, 116
[120] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. IEEE Transactions on pattern analysis and machine intelligence, 38(3):447–461, 2015. 97, 98
[121] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W Tsang. The emerging trends of multi-label learning. IEEE transactions on pattern analysis and machine intelligence, 44(11):7955–7974, 2021. 4
[122] Yang Liu, Hao Cheng, and Kun Zhang. Identifiability of label noise transition matrix. arXiv preprint arXiv:2202.02016, 2022. 104, 106
[123] Yang Liu, Hao Cheng, and Kun Zhang. Identifiability of label noise transition matrix. In International Conference on Machine Learning, pages 21475–21496. PMLR, 2023. 14
[124] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4258–4267, 2022. 67
[125] Yuyuan Liu, Yu Tian, Chong Wang, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Translation consistent semi-supervised segmentation for 3d medical images. arXiv preprint arXiv:2203.14523, 2022. 67
[126] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, 2016. 74

[127] David G Lowe. Object recognition from local scale-invariant features. In Proceedings of the seventh IEEE international conference on computer vision, volume 2, pages 1150–1157. Ieee, 1999. 68

[128] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In International Conference on Machine Learning, pages 6448–6458. PMLR, 2020. xv, 65, 67, 75, 76, 133

[129] Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In International Conference on Machine Learning, 2020. 107, 114

[130] Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 6500–6510. PMLR, 2020. 107, 109

[131] Congbo Ma et al. Multi-label thoracic disease image classification with cross-attention networks. In MICCAI, pages 730–738. Springer, 2019. 27, 56, 57, 58, 65, 66, 75, 76

[132] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In International Conference on Machine Learning, pages 6543–6553. PMLR, 2020. 3, 14, 67

[133] Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. Radiology, 294(2):421–431, 2020. viii, 57, 58, 74, 76

[134] Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". Advances in Neural Information Processing Systems, 30, 2017. xiii, 3, 58, 76, 86, 88, 97, 98, 106

[135] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. arXiv preprint arXiv:2007.07314, 2020. 6, 7, 53, 58, 81

[136] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE transactions on pattern analysis and machine intelligence, 41(8):1979–1993, 2018. 11, 34

[137] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? Advances in neural information processing systems, 32, 2019. 67, 68

[138] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj

Tewari. Learning with noisy labels. In Advances in neural information processing systems, pages 1196–1204, 2013. 2

[139] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. Nucleic acids research, 31(13):3812–3814, 2003. 1

[140] Nam Nguyen and Rich Caruana. Classification with partial labels. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 551–559, 2008. 2

[141] Peng Ni, Su-Yun Zhao, Zhi-Gang Dai, Hong Chen, and Cui-Ping Li. Partial label learning via conditional-label-aware disambiguation. Journal of Computer Science and Technology, 36(3):590–605, 2021. 4

[142] 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. 13

[143] Luke Oakden-Rayner. Exploring large-scale public medical image datasets. Academic radiology, 27(1):106–112, 2020. 15, 52, 65, 72, 81

[144] Luke Oakden-Rayner. Exploring large-scale public medical image datasets. Academic radiology, 27(1):106–112, 2020. 15, 52, 65

[145] Diego Ortego, Eric Arazo, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Multi-objective interpolation training for robustness to label noise. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6606–6615, 2021. 14, 106

[146] 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. 42

[147] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32, 2019. 25, 56, 74, 94

[148] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1944–1952, 2017. 3, 58, 76, 97, 98, 99, 104, 106, 113, 114, 116

[149] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. arXiv preprint arXiv:1906.05474, 2019. xiii, 65, 69, 70, 79, 80, 81

[150] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global

vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014. 80

[151] Congyu Qiao, Ning Xu, and Xin Geng. Decompositional generation process for instance-dependent partial label learning. In The Eleventh International Conference on Learning Representations, 2022. 4

[152] Pranav Rajpurkar et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225, 2017. 27, 56, 57, 58, 65, 66

[153] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. Advances in neural information processing systems, 28, 2015. 2

[154] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 82–91, 2021. 4, 65

[155] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. arXiv preprint arXiv:2101.06329, 2021. 2, 5, 12, 20

[156] Mamshad Nayeem Rizve et al. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In International Conference on Learning Representations, 2020. 32, 34, 35, 41, 42, 43

[157] Esther Rolf, Nikolay Malkin, Alexandros Graikos, Ana Jojic, Caleb Robinson, and Nebojsa Jojic. Resolving label uncertainty with implicit posterior models. arXiv preprint arXiv:2202.14000, 2022. 105, 109, 112, 113, 115, 117

[158] Olga Russakovsky et al. Imagenet large scale visual recognition challenge. IJCV, 115(3):211–252, 2015. 1, 56, 74

[159] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. Advances in neural information processing systems, 29, 2016. 8

[160] Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. Citeseer. 2

[161] Burr Settles. Active learning literature survey. 2009. 32

[162] Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In International Conference on Machine Learning, pages 5739–5748. PMLR, 2019. 3, 106

[163] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. Transductive semi-supervised deep learning using min-max features. In Proceedings of the European Conference on Computer Vision (ECCV), pages

299–315, 2018. 34

[164] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 94

[165] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. A co-regularization approach to semi-supervised learning with multiple views. In Proceedings of ICML workshop on learning with multiple views, volume 2005, pages 74–79. Citeseer, 2005. 87, 88

[166] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5972–5981, 2019. 2

[167] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In Computer Vision, IEEE International Conference on, volume 3, pages 1470–1470. IEEE Computer Society, 2003. 68

[168] Josef Sivic and Andrew Zisserman. Efficient visual search of videos cast as text retrieval. IEEE transactions on pattern analysis and machine intelligence, 31(4):591–606, 2008. 68

[169] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. Statistics and computing, 14:199–222, 2004. 1, 2

[170] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems, 33, 2020. 2, 5, 12, 32, 34

[171] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In International Conference on Machine Learning, pages 5907–5915. PMLR, 2019. 15, 16, 86, 93, 94, 99, 104, 112, 114, 115

[172] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. IEEE Transactions on Neural Networks and Learning Systems, 2022. 6

[173] Kaiwei Sun, Zijian Min, and Jin Wang. Pp-pll: Probability propagation for partial label learning. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II, pages 123–137. Springer, 2020. 4

[174] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2818–2826, 2016. 67

[175] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11662–11671, 2020. 53

[176] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification

by keeping the good and removing the bad momentum causal effect. In NeurIPS, 2020. 53

[177] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Advances in neural information processing systems, pages 1195–1204, 2017. 2, 5, 13, 20, 22, 23, 24, 26, 34, 42, 43, 44

[178] Yu Tian, Gabriel Maicas, Leonardo Zorron Cheng Tao Pu, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. 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. 20, 32

[179] 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. MICCAI2021, 2021. 32

[180] Yu Tian, Leonardo ZCT Pu, Rajvinder Singh, Alastair D Burt, and Gustavo Carneiro. One-stage five-class polyp detection and classification. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pages 70–73. IEEE, 2019. 20

[181] Yingjie Tian, Xiaotong Yu, and Saiji Fu. Partial label learning: Taxonomy, analysis and outlook. Neural Networks, 2023. 105, 107, 117

[182] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data, 5(1):1–9, 2018. xi, xiv, 16, 20, 24, 28, 32, 34, 35, 41, 119, 125, 126

[183] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. It takes (only) two: Adversarial generator-encoder networks. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018. 8

[184] Balagopal Unnikrishnan, Cuong Manh Nguyen, Shafa Balaram, Chuan Sheng Foo, and Pavitra Krishnaswamy. Semi-supervised classification of diagnostic radiographs with noteacher: A teacher that is not mean. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23, pages 624–634. Springer, 2020. 13, 20, 26, 35, 41, 43

[185] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(11), 2008. 80, 132, 135

[186] Deng-Bao Wang, Li Li, and Min-Ling Zhang. Adaptive graph guided disambiguation for partial label learning. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 83–91, 2019. 4

[187] Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and

Junbo Zhao. Pico: Contrastive label disambiguation for partial label learning. In International Conference on Learning Representations, 2021. 4

[188] Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao. Pico+: Contrastive label disambiguation for robust partial label learning, 2022. 106, 107, 109

[189] Qian Wang, Ning Jia, and Toby P Breckon. A baseline for multi-label image classification using an ensemble of deep convolutional neural networks. In 2019 IEEE International Conference on Image Processing (ICIP), pages 644–648. IEEE, 2019. 12, 46

[190] Wei Wang and Min-Ling Zhang. Partial label learning with discrimination augmentation. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 1920–1928, 2022. 4

[191] Xudong Wang, Long Lian, and Stella X Yu. Unsupervised selective labeling for more effective semi-supervised learning. In European Conference on Computer Vision, pages 427–445. Springer, 2022. 122

[192] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2097–2106, 2017. viii, x, xi, xiv, 15, 16, 20, 22, 24, 27, 28, 32, 34, 35, 41, 43, 44, 52, 53, 55, 56, 57, 86, 104, 119, 125, 126, 127

[193] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2097–2106, 2017. viii, x, 65, 66, 72, 74, 75, 76, 81, 131, 132, 134, 135

[194] Xudong Wang, Zhirong Wu, Long Lian, and Stella X Yu. Debiased learning from naturally imbalanced pseudo-labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14647–14657, 2022. 123

[195] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8688–8696, 2018. 14, 106

[196] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In Proceedings of the IEEE International Conference on Computer Vision, pages 322–330, 2019. 3, 14, 67

[197] Chen Wei, Lingxi Xie, Xutong Ren, Yingda Xia, Chi Su, Jiaying Liu, Qi Tian, and Alan L Yuille. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1910–1919, 2019. 13

[198] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13726–13735, 2020. xiii, 3, 13, 86, 88, 97, 98, 99, 100, 106, 115

[199] Jiaheng Wei, Hangyu Liu, Tongliang Liu, Gang Niu, and Yang Liu. Understanding (generalized) label smoothing whenlearning with noisy labels. arXiv preprint arXiv:2106.04149, 2021. 114

[200] Jiaheng Wei, Hangyu Liu, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Yang Liu. To smooth or not? when label smoothing meets noisy labels. Learning, 1(1):e1, 2021. xv, 65, 67, 68, 75, 76, 133

[201] Jiaheng Wei and Yang Liu. When optimizing f-divergence is robust with label noise. In International Conference on Learning Representation, 2021. 114

[202] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. arXiv preprint arXiv:2110.12088, 2021. 2, 15

[203] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022. ix, 112, 114, 143

[204] Pengxiang Wu, Songzhu Zheng, Mayank Goswami, Dimitris Metaxas, and Chao Chen. A topological filter for learning with label noise. Advances in neural information processing systems, 33:21382–21393, 2020. 65, 68, 70

[205] Xiaoxia Wu, Ethan Dyer, and Behnam Neyshabur. When do curricula work? In International Conference on Learning Representations, 2020. 6

[206] Xiaoxia Wu, Ethan Dyer, and Behnam Neyshabur. When do curricula work? In International Conference on Learning Representations, 2021. 34

[207] Zhi-Fan Wu, Tong Wei, Jianwen Jiang, Chaojie Mao, Mingqian Tang, and Yu-Feng Li. Ngc: a unified framework for learning with open-world noisy data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 62–71, 2021. 65, 68

[208] Xiaobo Xia, Jiankang Deng, Wei Bao, Yuxuan Du, Bo Han, Shiguang Shan, and Tongliang Liu. Holistic label correction for noisy multi-label classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1483–1493, 2023. 123

[209] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. Advances in Neural Information Processing Systems, 33:7597–7610, 2020. ix, 3, 14, 15, 65, 67, 93, 95, 97, 98, 99, 104, 106, 112, 113, 114, 116

[210] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? Advances in Neural Information Processing Systems, 32, 2019. 3, 6, 52, 97, 98, 114

[211] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2691–2699, 2015. 3, 15, 86, 93, 100, 104, 112

[212] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning with noisy label identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7):3676–3687, 2021. 68

[213] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. Advances in neural information processing systems, 33:6256–6268, 2020. 12

[214] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10687–10698, 2020. 12, 34

[215] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In Advances in Neural Information Processing Systems, pages 6225–6236, 2019. 113, 114

[216] Youjiang Xu, Linchao Zhu, Lu Jiang, and Yi Yang. Faster meta update strategy for noise-robust deep learning. In CVPR, pages 144–153, 2021. ix, 15, 94, 98, 99, 112, 114, 115, 143

[217] Cheng Xue, Lequan Yu, Pengfei Chen, Qi Dou, and Pheng-Ann Heng. Robust medical image classification from noisy labeled data with global and local representation guided co-training. IEEE Transactions on Medical Imaging, 2022. viii, 57, 58, 59, 74, 75, 76

[218] Shuo Yang, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, and Tongliang Liu. Estimating instance-dependent bayes-label transition matrix using a deep neural network. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of Proceedings of Machine Learning Research, pages 25302–25312. PMLR, 2022. 113, 114

[219] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. Advances in Neural Information Processing Systems, 33, 2020. 20

[220] Yu Yao, Tongliang Liu, Mingming Gong, Bo Han, Gang Niu, and Kun Zhang. Instance-dependent label-noise learning under a structural causal model.

Advances in Neural Information Processing Systems, 34:4409–4420, 2021. xiv, 8, 14, 15, 97, 98, 99, 104, 105, 106, 107, 109, 113, 114, 115, 116, 143

[221] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. arXiv preprint arXiv:2006.07805, 2020. 14, 65, 67

[222] Donggeun Yoo and In So Kweon. Learning loss for active learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 93–102, 2019. 2

[223] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine, 13(3):55–75, 2018. 86, 104

[224] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? arXiv preprint arXiv:1901.04215, 2019. xiii, 8, 13, 86, 88, 99, 100

[225] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In Proceedings of the IEEE/CVF international conference on computer vision, pages 1476–1485, 2019. 13, 20, 22

[226] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. Communications of the ACM, 64(3):107–115, 2021. 13, 87, 88

[227] Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. Delving deep into label smoothing. IEEE Transactions on Image Processing, 30:5984–5996, 2021. 65, 67, 75, 76

[228] Fei Zhang, Lei Feng, Bo Han, Tongliang Liu, Gang Niu, Tao Qin, and Masashi Sugiyama. Exploiting class activation value for partial-label learning. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022. 107, 109

[229] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017. 11, 39, 46, 97, 98, 99, 107, 115

[230] Min-Ling Zhang and Fei Yu. Solving the partial label learning problem: An instance-based approach. In IJCAI, pages 4048–4054, 2015. 4

[231] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. Pattern recognition, 40(7):2038–2048, 2007. 2

[232] Yang Zhang, Boqing Gong, and Mubarak Shah. Fast zero-shot image tagging. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5985–5994. IEEE, 2016. 65, 71

[233] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. arXiv preprint arXiv:2103.07756, 2021. 98, 99, 115

[234] Ziqi Zhang, Yuexiang Li, Hongxin Wei, Kai Ma, Tao Xu, and Yefeng Zheng. Alleviating noisy-label effects in image classification via probability transition matrix. arXiv preprint arXiv:2110.08866, 2021. 130

[235] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In Advances in neural information processing systems, pages 8778–8788, 2018. 3, 14, 67

[236] Wenting Zhao and Carla Gomes. Evaluating multi-label classifiers with noisy labels. arXiv preprint arXiv:2102.08427, 2021. 68

[237] Hong-Yu Zhou and et al. Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 398–407. Springer, 2020. 23, 24

[238] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16816–16825, 2022. 80

[239] Yu Zhou and Hong Gu. Geometric mean metric learning for partial label data. Neurocomputing, 275:394–402, 2018. 4

[240] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. National science review, 5(1):44–53, 2018. 2

[241] Zhaowei Zhu, Tongliang Liu, and Yang Liu. A second-order approach to learning with instance-dependent label noise. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10113–10123, 2021. 97, 98, 114