



THE UNIVERSITY
of ADELAIDE

Deep Learning for Scene Text
Detection, Recognition, and
Understanding

XINYU WANG

A thesis submitted for the degree of
DOCTOR OF PHILOSOPHY
The University of Adelaide

May 19, 2023

Contents

Abstract	xiii
Declaration of Authorship	xv
Acknowledgements	xvii
Publications	xix
1 Introduction	1
1.1 Background and Motivation	1
1.2 Contribution	5
1.3 Thesis Outline	6
2 Literature Review	7
2.1 Text Detection	7
2.2 Text Recognition	10
2.3 Text Spotting	13
2.4 Downstream OCR Applications	15
3 Benchmarking OCR systems: Datasets, Metrics, Methods	17
3.1 Introduction	19
3.1.1 Related Work	22
3.2 Unfair Comparison Between OCR Methods	24
3.2.1 Dataset Issues	24
3.2.2 Metric Issues	25
3.3 Unified OCR Benchmark	28
3.3.1 Data Collection	28
3.3.2 Evaluation Metric	30
3.3.3 Preliminary Experiments	31
3.3.3.1 Pre-training Matters	31
3.3.3.2 Scenario Matters	33
3.3.3.3 Language Matters	34
3.4 Experiments	34

3.4.1	Comparisons of State-of-the-art Methods	34
3.4.2	Discussion	36
3.5	Conclusion	37
4	Synthesizing High-Quality License Plates via a Text-to-Plate Network	39
4.1	Introduction	41
4.2	Related Work	43
4.2.1	License Plate Recognition	43
4.2.2	Synthetic Text Generation	45
4.2.3	Text-to-image Generation	46
4.3	Methods	46
4.3.1	Text to License Plate Network	46
4.3.1.1	Variational Auto-encoder	48
4.3.1.2	Transformer	50
4.3.2	A Simple Yet Strong Recognizer	52
4.4	Experiments	53
4.4.1	Datasets	53
4.4.2	Implementation Details	54
4.4.3	Comparison with real data	55
4.4.4	Comparison with other text synthesis approaches	56
4.4.5	Ablation Study	63
4.4.6	Results on CCPD-2018	64
4.4.7	Results on Extensive Benchmarks	65
4.4.7.1	CCPD-Green	65
4.4.7.2	CLPD	65
4.4.7.3	AOLP	65
4.5	Conclusion	66
5	On the General Value of Evidence, and Bilingual Scene-text Visual Question Answering	67
5.1	Introduction	70
5.1.1	Related Work	74
5.1.2	Text-based VQA	74
5.2	Proposed Dataset: EST-VQA	76
5.2.1	Data Collection	76
5.2.2	Evidence-based Evaluation (EvE) Metric	78
5.2.3	Tasks	78
5.3	Baselines and Results	82

5.3.1	Baseline Methods	82
5.3.2	Results	84
5.4	Conclusion	86
5.5	Supplementary	86
5.5.1	Annotation Guidelines	86
5.5.2	Annotation Pipeline	87
5.6	More Annotation Examples	89
5.7	More Examples of Unreasonable Output in Conventional Text-VQA dataset	92
6	Conclusions	95
6.1	Future Work	96

List of Figures

1.1	Application of OCR technology in various scenarios: (a) document analysis; (b) license plate; (c) handwritten mathematical expression; (d) text editing; (e) natural scene; (f) text visual question answering.	2
1.2	(a) Text Detection: identifying and locating text in the image. (b) Text Recognition: recognizing the cropped image patches of detected texts. (c) End-to-end Spotting: simultaneously detecting and recognizing texts in images.	3
1.3	A visual representation of the various scenarios in the OCR task. The figure shows examples of different types of OCR scenarios, including natural scene text, document, handwritten text, etc.	4
2.1	The text detection task is defined as the process of localizing the text region in an input image.	7
2.2	The evolution of annotation forms in text detection datasets over the years. (a) shows an example of annotation with horizontal rectangles, which is the most traditional form used in generic object detection. (b) illustrates the use of rotated rectangles, which is able to capture the multi-oriented text in images. (c) and (d) show the annotation forms of quadrilaterals and polygons, respectively, which can effectively handle texts of arbitrary shapes.	9
2.3	Novel representations proposed by recent works that are designed for arbitrary text detection.	10
2.4	The text recognition task is defined as the process of converting the text in images to machine-readable formats.	10
2.5	Two mainstream techniques used in text recognition.	11
2.6	Text spotting, also known as end-to-end text detection and recognition, aims to simultaneously detect and recognize texts in images.	13
2.7	Example methods of two mainstream pipelines of text spotters. (a) is a typical regression-based method ABCNet [102]. (b) is a popular segmentation-based framework MaskTextSpotter [91].	14
2.8	SPTS [129]	15

3.1	Comparison of typical evaluation protocols between object detection and OCR. The former usually uses definite experimental settings, including train/test datasets, data augmentation pipelines, training schedules, etc., to ensure fair comparisons, while the latter uses many indefinite configurations.	21
3.2	All of these predictions are treated as the same true positive under the VOC metric; thus, the detection performance is usually overestimated. Green and red bounding boxes represent GT and prediction, respectively.	26
3.3	In most OCR benchmarks, the prediction contributes to the accuracy only if it is exactly the same as the ground truth, which sometimes underestimates the performance of recognizers. . . .	26
3.4	Data distribution of the UniOCR benchmark from the aspect of (a) Scenario; (b) Language; (c) Annotation Granularity; and (d) Bounding Box Form. It should be noted that as one image may contain multiple annotation granularity or bounding box forms, <i>e.g.</i> , Latin instances are labeled in word-level granularity while Chinese instances followed line-level labeling rules, the summations of percent values in (c) and (d) are larger than 1.	29
3.5	Detected boxes with different overlaps with GT.	30
3.6	Performance of models using different pre-training strategies. Solid and dashed lines represent end-to-end and detection-only AP, respectively.	32
4.1	Examples of license plates from the CCPD dataset [182]. (a) CCPD-2018 contains 300k blue license plates, and each consists of 1 province code, 1 city code, and 5-char ID. (b) CCPD dataset contains 10k green license plates, and each consists of 1 province code, 1 city code, and 6-char ID.	42
4.2	Data distribution of the CCPD-2018 [182] training split shows the majority ($\sim 90\%$) of car license plate samples are collected from a single city, which has identical province and city code (皖A).	43
4.3	Structure of the Text to License Plate Network (TLPNet). The TLPNet consists of three parts: 1) a dVAE that compresses images into tokens; 2) an embedding layer that encodes each input license plate string into textual features; 3) a transformer that aggressively models the combined image and text features. . . .	44
4.4	License plate number embedding.	50

4.5	The network structure of the baseline recognizer, is composed of three: spatial transformer, backbone network, and bidirectional LSTM.	51
4.6	Comparison of real images and synthetic data generated by different algorithms.	54
4.7	Synthetic LPs that are generated by our methods. Compared to the existing methods that directly render digits and characters on a blank template, the proposed methods learn from existing data to generate more realistic photos. First row: generated LP with less tilt and rotation. Second row: generated LP with large tilt.	62
4.8	Failure cases of generated LPs. As the available training data of these provinces is very limited (<i>e.g.</i> 藏 (7); 青 (8); 云 (10); 甘 (10)), the proposed method failed to generate high resolution province code.	63
4.9	Examples of the incorrect recognition results. The samples from the three rows are from CCPD-2018, CCPD-Green, and CLPD, respectively.	63
5.1	Requiring that vision-and-language methods provide evidence for their decisions encourages the development of approaches that depend on reasoning and thus that are better able to generalize to new situations. It also helps to build up confidence in the provided answer.	71
5.2	Some example images and QA pairs from the Text-VQA proposed in [151]. Four different types of issues are shown. (a) questions that can be answered without reading image text; (b) questions that have more than one correct answer; (c) questions that require a large amount of external knowledge to answer; (d) questions that require skills that cannot be learned from the training data alone.	72
5.3	A comparison of conventional (LoRRA [151]), and evidence-based VQA methods.	72
5.4	Illustration of the mainstream VQA models. D_q , D_i , D_o and D_h are the dimensions of the word embedding, image feature, OCR token embedding and hidden vector representations respectively. N , N' and P indicate question length, number of OCR tokens and answer space. Blocks with dashed lines are optional modules used for text-based VQA.	74

5.5	Distribution of first four words in question sets of EST-VQA. . .	75
5.6	Percentage of question and answer length in EST-VQA dataset. Questions are tokenized by words. En and Ch stand for English and Chinese respectively.	77
5.7	In EvE metric, evidence in the form of bounding box should be provided as well as the predicted answer. Green and red bounding boxes are ground-truth and predicted evidence respectively. Incorrect: (a) answer without evidence; (b) answer with inappropriate evidence; (c) answer with insufficient evidence. Correct: (d) answer with appropriate evidence. It is worth mentioning that all of the above answers would be marked as correct in the conventional VQA evaluation metric because all of them give the right answer ‘2’.	79
5.8	Overview of the QA R-CNN architecture.	79
5.9	CLC score under different τ	83
5.10	Visualization of the output answers on the EST-VQA dataset from different models (first four images). Green and Red bounding boxes are ground-truth and predicted evidence by QA R-CNN. (More examples can be found at https://arxiv.org/abs/2002.10215)	84
5.11	Labelling Tool. At the first stage, annotators are asked to label a rectangle or quadrilateral bounding-box for a potential answer.	89
5.12	Two bounding-box labelling modes are available in the annotation tool. Annotators are asked to select the most appropriate one by considering the tightness between the text and bounding box.	89
5.13	Labelling Tool. In the second stage, annotators are asked to come up with a question based on the corresponding text covered by the bounding box.	90
5.14	Word cloud of majority answers in STE-VQA dataset.	90
5.15	Examples of English questions.	91
5.16	Examples of Chinese questions.	91
5.17	Unreasonable Output Part A	92
5.18	Unreasonable Output Part B	93

List of Tables

3.1	Accuracy on a typical scene text dataset Total-Text [24] reported by recent proposed text spotters. Almost all of these OCR methods used external data to train the model within different training settings; This means that the comparisons between the reported performance on such benchmarks may not reflect the actual effectiveness of the proposed OCR algorithms. Datasets: SynT (SynthText) [49], IC13 [70], IC15 [71], ArT [25], TT (Total-Text) [24], SCUT [203], MLT17 [123], MLT19 [124], CT (COCO-Text) [38], Pvt. represents private data. Star* means that only part of the data is used. Dagger† means that character-level supervision is included at the training stage.	20
3.2	Performance of ABCNet [102] on Total-Text [24] using different training data. Datasets: SynT [49], MLT [124], and TT [24]. . .	25
3.3	Cross validation of ABCNet [102] on Total-Text [24].	25
3.4	Evaluation metrics used in some OCR datasets.	26
3.5	Performance of ABCNet [102] on Total-Text [24] using different testing size. Official code uses (1000, 1824).	27
3.6	Performance of ABCNet [102] on Total-Text [24] and SCUT-CTW1500 [100] using different recognition resolution. Official code uses (8, 128) and (8, 32) for CTW and TT, respectively. . .	27
3.7	UniOCR covers 25 publicly available OCR datasets. Granu. represents annotation granularity, including Char (C), Word (W), and Line (L). Licenses are included in supplementary.	28
3.8	Data volume of UniOCR.	29
3.9	Precisions under Total-Text [24] and UniOCR metrics.	30
3.10	Detection AP of models trained with or without natural scene samples on subsets in different scenarios.	33
3.11	Detection AP of models trained with or without English text on subsets in different languages.	33
3.12	Performance comparison between state-of-the-art OCR methods on UniOCR in terms of H-mean score.	35

3.13	Performance on TotalText [24] without lexicon.	36
4.1	Network structure of the discrete variational auto-encoder. . . .	48
4.2	Performance comparison between models that are trained on real images and synthetic data on CCPD2018 test set. Minor* is a subset that excludes samples started with 皖.	56
4.3	Comparison of recognition performance using ground-truth and detection bounding box.	56
4.4	Comparison of performance using different data synthesis methods.	57
4.5	Performance comparison with state-of-the-art LP recognition methods on the CCPD-2018 test splits. Note that, for ‘Ours (Synthetic Data Only)’, only 6000 real labeled images are used for training the generator. For Ours ‘(Synthetic+Real Data)’, 6000 real labeled images for training the generator and standard CCPD-2018 training split for training the recognition network.	60
4.6	Impact of the number of labeled data while training the generator.	61
4.7	Performance of the proposed methods on the CCPD dataset. . .	61
4.8	Comparison of performance between ours and the state-of-the-art methods on the CLPD [196] dataset.	62
4.9	Comparison of performance between ours and the state-of-the-art methods on the AOLP [196] dataset.	62
5.1	A comparison of the amount and source of images between different text-based VQA datasets. #I and #Q indicate the number of images and questions respectively.	75
5.2	Volume of the EST-VQA dataset.	77
5.3	Quantitative results of the three tasks in EST-VQA dataset. Mono. and Bi. represent monolingual and bilingual model respectively while S and L are short (one word) and long (more than one word) answers. Scores in bold are the best performance across models.	81
5.4	Distribution of question number per image.	88
5.5	Distribution of answer type in the STE-VQA dataset.	89
5.6	A summary and comparison of different aspects between Text-VQA, ST-VQA and the proposed STE-VQA.	94

The University of Adelaide

Abstract

Deep Learning for Scene Text Detection, Recognition, and Understanding

by XINYU WANG

Detecting and recognizing texts in images is a long-standing task in computer vision. The goal of this task is to extract textual information from images and videos, such as recognizing license plates. Despite that the great progresses have been made in recent years, it still remains challenging due to the wide range of variations in text appearance. In this thesis, we aim to review the existing issues that hinder current Optical Character Recognition (OCR) development and explore potential solutions. Specifically, we first investigate the phenomenon of unfair comparisons between different OCR algorithms caused due to the lack of a consistent evaluation framework. Such an absence of a unified evaluation protocol leads to inconsistent and unreliable results, making it difficult to compare and improve upon existing methods. To tackle this issue, we design a new evaluation framework from the aspect of datasets, metrics, and models, enabling consistent and fair comparisons between OCR systems. Another issue existing in the field is the imbalanced distribution of training samples. In particular, the sample distribution largely depended on where and how the data was collected, and the resulting data bias may lead to poor performance and low generalizability on under-represented classes. To address this problem, we took the driving license plate recognition task as an example and proposed a text-to-image model that is able to synthesize photo-realistic text samples. By using this model, we synthesized more than one million samples to augment the training dataset, significantly improving the generalization capability of OCR models. Additionally, this thesis also explores the application of text vision question answering, which is a new and emerging research topic among the OCR community. This task challenges the OCR models to understand the relationships between the text and backgrounds and to answer the given questions. In this thesis, we propose to investigate evidence-based text VQA, which involves designing models that can provide reasonable evidence for their predictions, thus improving the generalization ability.

Declaration of Authorship

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.

I acknowledge the support I have received for my research through the provision of an Adelaide Graduate Research Scholarship.

Xinyu Wang

January 24, 2023

Acknowledgements

First and foremost, I would like to thank my Ph.D. supervisor, Prof. Chunhua Shen, for all his help during my study at the University of Adelaide. It is fortunate for me to work with such a researcher who is always enthusiastic about his work and willing to provide help and guidance whenever needed. Prof. Shen has always been a constant source of support, guidance, and inspiration for me throughout my Ph.D. journey.

I am also grateful to my co-authors, lab mates, and friends, without whom I could never have made it through those tough days. They were always there to lend a listening ear, offer a helping hand, and provide words of encouragement when I needed it the most.

Finally, I am forever grateful to my parents. They helped me to not only survive but also to thrive, and for that, I will always be thankful.

Publications

The following are articles that have been published or submitted for publication during my Ph.D. study, with those marked with an * being included in this thesis:

- * On the General Value of Evidence, and Bilingual Scene-text Visual Question Answering
Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jing, Chee Seng Chan, Anton van den Hengel
 The Conference on Computer Vision and Pattern Recognition (CVPR), 2020
- * Benchmarking OCR systems: Datasets, Metrics, Methods
Xinyu Wang, Yuliang Liu, Chunhua Shen
 Submitted to Pattern Recognition (PR).
- * Synthesizing High-Quality License Plates via a Text-to-Plate Network
Xinyu Wang, Yuliang Liu, Chunhua Shen Submitted to IEEE Transactions on Circuits and Systems on Video Technology (T-CSVT).
- SPTS: Single-point Text Spotting
 Dezhi Peng, **Xinyu Wang**, Yuliang Liu, Jiaxin Zhang, Mingxin Huang, Songxuan Lai, Jing Li, Shenggao Zhu, Dahua Lin, Chunhua Shen, Xiang Bai, Lianwen Jin
 The ACM International Conference on Multimedia (ACMMM), 2022
- Improving Handwritten Mathematical Expression Recognition via Similar Symbol Distinguishing
 Zhe Li, **Xinyu Wang**, Yuliang Liu, Lianwen Jin, Yichao Huang, and Kai Ding
 IEEE Transactions on Multimedia (T-MM), Accepted.
- Exploring the Capacity of an Orderless Box Discretization Network for Multi-orientation Scene Text Detection
 Yuliang Liu, Tong He, Hao Chen, **Xinyu Wang**, Canjie Luo, Shuaitao Zhang, Chunhua Shen, Lianwen Jin
 International Journal of Computer Vision (IJCV), 2021
- ICDAR 2021 Competition on Integrated Circuit Text Spotting and Aesthetic Assessment

Chun Chet Ng, Akmalul Khairi Bin Nazaruddin, Yeong Khang Lee, **Xinyu Wang**, Yuliang Liu, Chee Seng Chan, Lianwen Jin, Yipeng Sun, Lixin Fan
The International Conference on Document Analysis and Recognition (ICDAR), 2021

- When IC Meets Text: Towards a Rich Annotated Integrated Circuit Text Dataset
Chun Chet Ng, Che-Tsung Lin, Zhi Qin Tan, **Xinyu Wang**, Jie Long Kew, Chee Seng Chan, Christopher Zach
Submitted to Pattern Recognition (PR).
- SPTS v2: Single-Point Scene Text Spotting
Yuliang Liu, Jiaxin Zhang, Dezhi Peng, Mingxin Huang, **Xinyu Wang**, Jingqun Tang, Can Huang, Dahua Lin, Chunhua Shen, Xiang Bai, Lianwen Jin
arXiv 2023, Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI).

Chapter 1

Introduction

Texts are tools used by humans to write and record information, and their origins can be traced back to as early as the third millennium B.C. Throughout human history, texts have evolved and played a crucial role in the development of civilization. From the earliest forms of hieroglyphics in ancient Egypt, to the creation of the Greek alphabet in the fourth century B.C., and the widespread use of Latin script during the Middle Ages, texts have provided a means for humans to communicate, record, and preserve information. Today, texts continue to be an essential aspect of modern society, used in a variety of contexts, including education, business, and everyday communication. As a result, the automation of text detection and recognition in images and videos has been a topic of significant research interest for scholars within the computer vision community over the past decades [61, 70, 71, 107, 111, 125, 129, 143, 165, 169, 185].

1.1 Background and Motivation

Text detection and recognition, also known as Optical Character Recognition (OCR), is a technology that enables the conversion of images with text into machine-readable form. This process allows for the automatic processing and analysis of text contained in images and documents, which has been widely adopted in many real-world applications (see Figure 1.1), including document analysis [17, 56], car license recognition [5, 6, 15, 18, 31], image retrieval [37, 68, 112, 164], handwritten text recognition [33, 120, 128, 132, 190], text generation and manipulation [59, 78, 122, 176, 199], and scene text spotting [85, 97, 102, 104, 109]. In addition, it is interesting to note that OCR tasks have also been combined with Natural Language Processing (NLP) techniques in recent years due to the rich semantic information contained in the text. This convergence has resulted in a new realm of cross-modal research, which allows for a deeper understanding of the context and semantics of the text, leading to improved performance in traditional OCR tasks. Furthermore, such a combination has

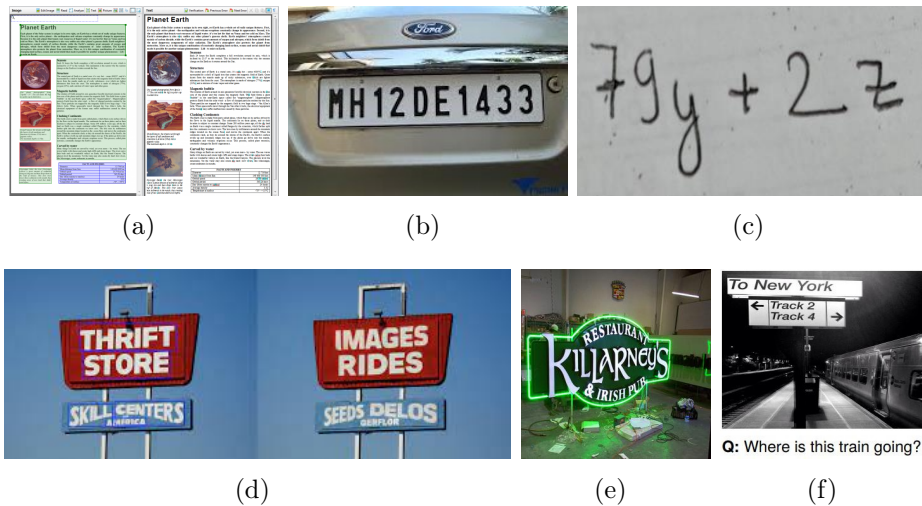


FIGURE 1.1. Application of OCR technology in various scenarios: (a) document analysis; (b) license plate; (c) handwritten mathematical expression; (d) text editing; (e) natural scene; (f) text visual question answering.

sparked the emergence of a series of novel tasks that merge both CV and NLP techniques, such as text-based visual question answering (VQA) [62, 116, 118, 151, 173] and document understanding [48, 206]. These tasks necessitate the incorporation of language processing abilities in order to accurately interpret and respond to queries or understand the content of documents, which has opened up exciting new opportunities for advancing the capabilities of OCR-based systems.

OCR is typically divided into two phases: text detection (Figure 1.2(a)) and text recognition (Figure 1.2(b)). The former involves identifying and locating written or printed text within an image or video, while the latter entails converting the identified text into a format that can be interpreted by a machine, such as ASCII or Unicode. These stages are crucial for the successful completion of an OCR task as they work in tandem to extract and decipher the text within the input media. In recent years, end-to-end text spotting (Figure 1.2(c)) models have become increasingly popular, as they can simultaneously perform both text detection and recognition. Compared to the traditional approaches which typically involve separate stages, end-to-end methods are able to make use of the context and relationship between the detected text and surrounding regions to improve the accuracy of the recognition process. They also tend to be more efficient, as they do not require the execution of multiple stages or the use of complex pipelines, such as cropping detected texts into patches.

The history of OCR can be dated back to the early 1900s when the basic

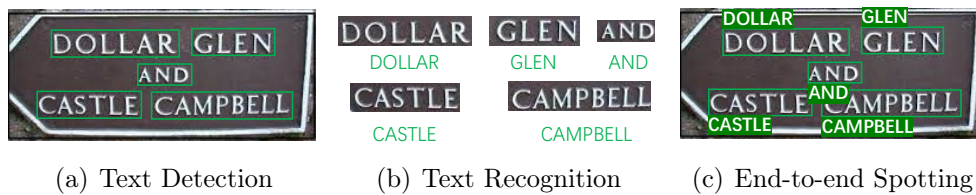


FIGURE 1.2. (a) Text Detection: identifying and locating text in the image. (b) Text Recognition: recognizing the cropped image patches of detected texts. (c) End-to-end Spotting: simultaneously detecting and recognizing texts in images.

concept of modern OCR technology was first proposed by German engineer Tauschek [159]. Tauschek recognized the potential for automated recognition of handwritten text in scanned documents, which would greatly improve the efficiency and accuracy of document processing and analysis. His proposal marked the beginning of the development of OCR technology, which has since undergone significant evolution and improvement. After nearly a century of development, and especially in recent decades with the advent of deep learning, OCR technology has become an essential tool in various fields and has enabled the automation of a wide range of tasks involving text recognition and analysis.

Despite its widespread adoption and success in many commercial and industrial applications, OCR technology still faces many challenges and difficulties.

- One of the main challenges is the variability and complexity of text in real-world scenarios (see Figure 1.3), which can greatly impact the performance of OCR systems. For example, text can appear in various fonts, sizes, colors, and languages, and can be distorted, degraded, or occluded by various factors, such as noise, blur, and shadows. These variations can significantly affect the ability of OCR systems to accurately detect and recognize text and can lead to errors or inconsistencies in the output.
- The second challenge is the limited robustness and generalization capabilities, which can limit the OCR systems' performance on unseen or novel data. For example, many OCR models are designed specifically for certain types of text, such as handwritten text or natural scene text, and may not be effective for other scenarios. Such a lack of generalizability can limit the flexibility and adaptability of OCR systems and can hinder their performance when applied to different data and domains.
- Another challenge is the lack of a standardized evaluation framework. This can make it difficult to objectively compare the performance of different OCR algorithms, as different studies may use different training data,

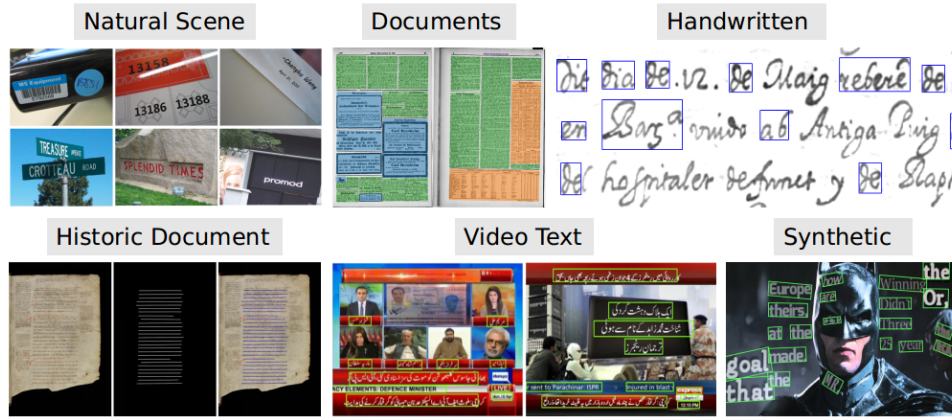


FIGURE 1.3. A visual representation of the various scenarios in the OCR task. The figure shows examples of different types of OCR scenarios, including natural scene text, document, handwritten text, etc.

optimization strategies, and testing parameters. This lack of uniformity can result in inconsistent and inconsistent comparisons between models, hindering the development and advancement of OCR research. It can be challenging to assess the strengths and weaknesses of different algorithms and to identify the most promising directions for future research when the evaluation criteria are not consistent.

- The last challenge is the lack of tight integration between OCR technology and downstream tasks, such as text-based visual question answering and document understanding. In many cases, these tasks simply use off-the-shelf OCR models to extract text from images. Such disconnection may lead to sub-optimal performance, as the OCR model may not be tailored to the specific requirements of the downstream tasks. Thus the two models may not work in harmony to effectively extract and utilize the relevant information from the input data.

Overall, the above mentioned difficulties faced by OCR technology highlight the need for further research and development. In this thesis, we aim to address some of the above challenges. For example, we design a unified OCR benchmark framework that can evaluate the performance of OCR models on a variety of text types and scenarios and provide a fair comparison between different algorithms. We also propose a text synthesis model for generating photo-realistic samples that can augment and balance the training dataset for license plate recognition. Additionally, we explore the integration of OCR technology with a downstream task, text visual question answering. By incorporating OCR models into text VQA, we demonstrate how this integration can enhance the overall reasoning

abilities of the model and improve its generalization capabilities. Our approach leverages the strengths of both OCR and text VQA models and combines them in a way that allows them to work seamlessly and cohesively to extract and utilize relevant information from the input data.

1.2 Contribution

The main focus of this thesis is to explore the methods of addressing existing challenges in the field of OCR, as outlined below:

- The absence of a standardized benchmark system within the OCR field results in inconsistent comparisons between state-of-the-art models, hindering an accurate assessment of their performance. To address this challenge, we design a new benchmark system that can be employed to objectively evaluate the performance of OCR models. Our designed approach encompasses three key elements: datasets, models, and metrics. Specifically, we unified the training and testing dataset, as well as the hyperparameters and evaluation metrics, to ensure fair and consistent comparisons across different models.
- Another challenge faced in the OCR task is the issue of imbalanced data. This refers to the fact that certain classes or categories within the data may have a disproportionately larger number of examples than others, leading to a biased model that performs inadequately on under-represented classes. To tackle this challenge, we propose to design a new text-to-image synthesis model that generates photo-realistic text samples specifically tailored to the task of license plate recognition. Such a model enables the augmentation of high-quality and diverse training samples for the OCR model, which in turn improves its generalization ability and robustness to different variations of license plate text. In addition, by generating a vast number of realistic samples (exceeding one million), we demonstrate that OCR models trained solely on synthetic data can also achieve comparable performance to those trained on real data.
- Beyond simply recognizing texts in images, it is also important for models to comprehend the textual contents. This poses a significant challenge as it requires a model to have a profound understanding of both visual and textual information. Specifically, we delve into the task of text visual question answering, which refers to the ability of a model to understand

and answer questions based on texts present in an image. One of the long-standing issues in this task is the lack of reasoning ability in current models. This means that the VQA models predict answers without providing any reasoning or explanation, resulting in poor generalizability. To address this challenge, we propose to investigate evidence-based text visual question answering, which involves designing models that can provide reasoning and evidence for their predictions, thus improving their generalization ability and robustness to unseen examples. Furthermore, we introduce a new dataset, as well as a novel metric, to facilitate the quantitative evaluation of model reasoning capability.

1.3 Thesis Outline

Based on the aforementioned contributions, we organize the thesis structure as follows:

Chapter 2 reviews the existing literature in the field of OCR, highlighting the development of state-of-the-art models over the past decades.

Chapter 3 examines the inconsistent comparisons between recently proposed OCR models and describes the design and implementation of the proposed new benchmark system for OCR models. It includes the datasets, models, and metrics used to ensure fair and consistent comparisons across different models.

Chapter 4 presents our work on license plate recognition. Specifically, we design a text-to-image synthesis model to generate photo-realistic samples, addressing the issue of imbalanced data. By generating more than one million synthetic samples, we demonstrate that a lightweight recognizer can achieve comparable performance by solely using these generated data.

Chapter 5 investigates the reasoning ability of the OCR model in the task of text visual question answering. We propose to quantitatively evaluate the reasoning ability of text VQA models by designing an evidence-based VQA system accompanied by both new datasets and models.

Chapter 6 summarizes the main contributions of this thesis and discusses future directions.

Chapter 2

Literature Review

The field of Optical Character Recognition (OCR) has undergone significant advancements in recent years, largely due to the advancements in deep learning technologies [106]. With the rise of deep neural networks, a wide range of models, datasets, and techniques have been proposed to improve the accuracy and efficiency of text detection and recognition in images. This chapter aims to provide an overview of the state-of-the-art OCR models and techniques across four key aspects: text detection, text recognition, text spotting, and downstream OCR applications. We will delve into the strengths and limitations of various models and techniques proposed in these areas, as well as the challenges that still exist in the field and the directions to which research is currently headed. Specifically, we will examine how deep learning techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been utilized to improve recognition performance in these tasks.

2.1 Text Detection

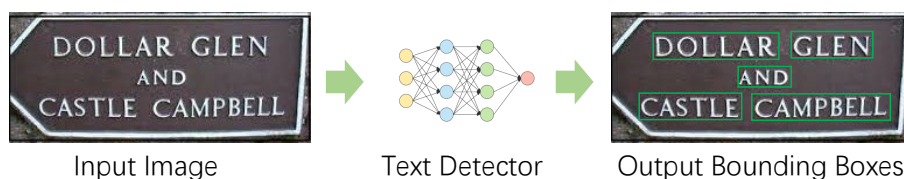


FIGURE 2.1. The text detection task is defined as the process of localizing the text region in an input image.

In the OCR task (See Figure 2.1), text detection is the first and crucial step in recognizing text in images. The text detection task is defined as the process of identifying the regions of an image that contain texts and then extracting the coordinates of these regions. Since its similarity with the generic object detection task, early text detection methods [27, 54, 64, 90, 98, 145, 198, 204] often borrowed ideas heavily from object detectors [36, 52, 95, 141]. Typically,

they use CNNs to extract deep features and a classifier to classify Regions of Interest (RoIs) generated by a Region Proposal Network (RPN). However, these methods often took into account the differences between text instances and general objects, such as aspect ratio and tilted angles, by redesigning the rules for generating RoIs.

For example, the Deep Matching Prior Network (DMPN) proposed by Liu et al. [98] introduced a novel approach to text detection by utilizing a quadrilateral sliding window. This approach effectively addresses the challenge of tilted text instances, which traditional horizontal sliding window methods used in generic object detectors struggle with. The use of this flexible quadrilateral sliding window allows DMPN to achieve a higher detection accuracy on multi-oriented text instances. Similarly, Jiang et al. [64] proposed a rotational region CNN called R²CNN. As shown in Figure ??, R²CNN modified the Fast R-CNN [35] model to classify text regions, and three different sizes of ROI Poolings, i.e., (7×7) , (11×3) , and (3×11) were employed for maximizing the text characteristics. Although these models derived from R-CNN series methods achieved promising performance on horizontal and multi-oriented text, they fail to handle text instances of arbitrary shapes, particularly curved text.

Therefore, there is a growing interest in developing novel text detection methods that can effectively handle texts of arbitrary shapes, highlighting the evolution of text detection datasets and annotation forms (see Figure 2.2). Initially, text detection datasets were mostly annotated with horizontal rectangles [17, 69, 70, 107] just as the form used in generic object detection. Later, to facilitate the detection of multi-oriented and arbitrarily shaped texts, text datasets began to include annotations for rotated rectangles [184], and even more complex shapes such as quadrilaterals [147, 188, 191] and polygons [24, 26, 100]. Such a shift in annotation forms reflects the increasing demand for text detection methods that can handle texts of arbitrary shapes, encouraging the emergence of novel text detectors [110, 158, 162, 163, 208] that can handle such complexities.

For example, Wang et al. [163] introduced TextRay, a novel text detection method that is capable of detecting texts of arbitrary shapes. TextRay utilizes top-down contour-based geometric parameters within a single-shot, anchor-free framework. As shown in Figure 2.3(a), it encodes complex geometric layouts into unified representations by utilizing a polar system and a bidirectional mapping scheme between shape space and parameter space. Zhu et al. [208] proposed a novel method called Fourier Contour Embedding (FCE). FCE models text instances in the Fourier domain, which enables the representation of text



(a) IC13 [70] (Horizontal Rect)



(b) TD500 [184] (Rotated Rect)



(c) NEOCR [121] (Quadrilaterals)



(d) Total Text [24] (Polygon)

FIGURE 2.2. The evolution of annotation forms in text detection datasets over the years. (a) shows an example of annotation with horizontal rectangles, which is the most traditional form used in generic object detection. (b) illustrates the use of rotated rectangles, which is able to capture the multi-oriented text in images. (c) and (d) show the annotation forms of quadrilaterals and polygons, respectively, which can effectively handle texts of arbitrary shapes.

contours as compact signatures, allowing for the efficient detection of texts with complex shapes and orientations. In addition, Tang [158] et al. handled curved and dense texts via a segmentation manner. They introduced a novel network for detecting dense and irregularly shaped text via Instance-aware Component Grouping (ICG), a bottom-up approach that offers a high degree of flexibility (see Figure 2.3(c)).

In summary, text detection methods have undergone significant evolution over the past decade, while early text detectors often borrowed heavily from generic object detection but failed to handle texts of arbitrary shapes. With the increasing demand for detecting multi-oriented and arbitrarily shaped texts, the annotation forms of text detection datasets have been shifted from horizontal rectangles to more complex shapes such as rotated rectangles, quadrilaterals,

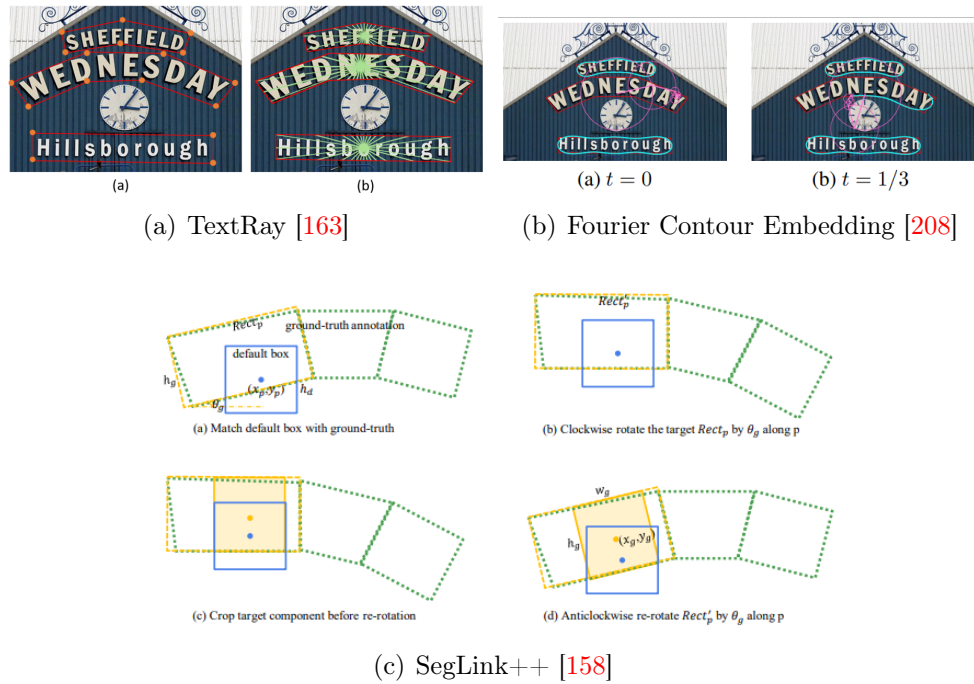


FIGURE 2.3. Novel representations proposed by recent works that are designed for arbitrary text detection.

and polygons, which encourages the emergence of novel text detectors. These methods have shown promising performance on various datasets, highlighting the potential for further advancements in scene text detection.

2.2 Text Recognition

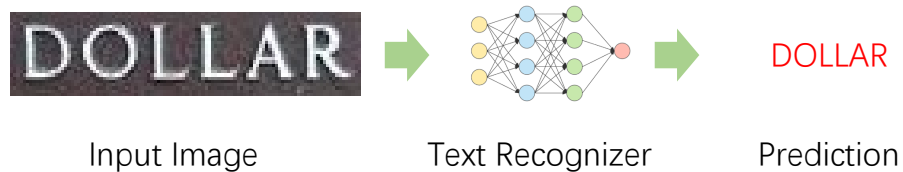


FIGURE 2.4. The text recognition task is defined as the process of converting the text in images to machine-readable formats.

As shown in Figure 2.4, text recognition is the task of converting cropped images of detected text (usually containing only a single word or sentence) into machine-readable formats, such as ASCII and Unicode. Technically, there are

two mainstream frameworks in this field, i.e., Connectionist Temporal Classification (CTC) [44] (see Figure 2.5(a)) and the encoder-decoder structure [157] (see Figure 2.5(b)). The former is a framework that focuses on mapping input sequences to output sequences while preserving the temporal structure of the input. On the other hand, the encoder-decoder structure utilizes a neural network to first encode the input image sequence to deep features, then decode the features to produce the output sequence.

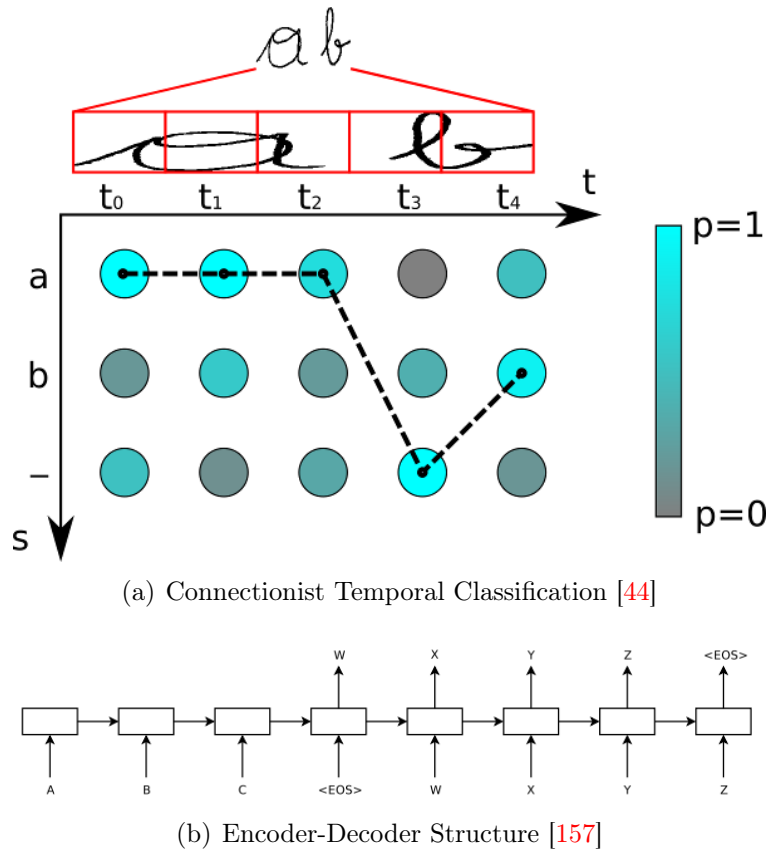


FIGURE 2.5. Two mainstream techniques used in text recognition.

The CTC decoding module was originally developed for speech recognition and was first adopted for handwritten text recognition by Graves et al. [45]. Different from speech recognition tasks where data is sequential in the time domain, the input image in text recognition can be viewed as a sequence of vertical pixel frames. The network produces a prediction for each frame, indicating the likelihood of different label types. The CTC rule is then applied to transform these predictions into a text string. During the training process, the loss is calculated by taking into account all possible per-frame predictions that can generate the target sequence through the use of CTC rules. This approach allows

for an end-to-end training process using only word-level annotations, eliminating the need for character-level annotations, which has now been widely used in text recognition task [96, 143, 153, 187]. For example, Shi [143] proposed a fully convolutional text recognizer called CRNN. The CRNN starts with a series of convolutional layers that extract a feature sequence from cropped image patches. These features are then passed through a deep bidirectional LSTM, which makes predictions for each frame of the sequence. Furthermore, the final step involves using a transcription layer, i.e., the CTC module, to convert the per-frame predictions from the LSTM into a label sequence.

The other widely used technique, i.e., encoder-decoder framework, was first proposed by Sutskever et al. [157] for the purpose of machine translation. In such a pipeline, the encoder takes in an input sequence and passes on its final hidden state to the decoder. The decoder then generates output in an auto-regressive manner. One of the key benefits of this framework is its ability to produce outputs of varying lengths, making it well-suited for tasks such as text recognition. Additionally, the encoder-decoder structure is often paired with the attention mechanism, allowing for the simultaneous alignment of inputs and outputs [12, 21, 80, 105]. For example, Bai et al. [12] presented a novel approach known as the edit probability (EP) for recognizing texts. The EP method endeavors to precisely calculate the chance of creating a string from the output series of a probability distribution based on the input image, taking into account the potential presence of omitted or extra characters. The benefit of this approach is that the training process can concentrate on rectifying omitted, extra, and unidentifiable characters, thereby reducing or eliminating the impact of misalignment issues.

However, although both CTC and encoder-decoder structures achieve promising performance on horizontal text, an inevitable issue arises when it comes to recognizing irregular text. This is because characters in the oriented and curved text are distributed over a 2-dimensional space, which makes it difficult to effectively represent them in feature spaces that are compatible with the CTC and encoder-decoder structures, as they are designed for 1D sequences. Therefore, directly compressing the features of oriented and curved text into 1D form can result in the loss of relevant information, leading to a decrease in recognition accuracy. To tackle this issue, a widely adopted solution is to utilize a rectification module [60, 144, 148, 183, 192] to transform the irregular inputs to a more canonical one. For example, Shi et al. [148] introduced a recognizer with flexible rectification called ASTER which consists of two main components: a rectification network and a recognition network. The rectification network is

inspired by the Spatial Transformer Network (STN) [60], straightening out the input text images. Specifically, the rectification is achieved through the use of a flexible Thin-Plate Spline transformation, which can handle a wide range of text irregularities and is trained without the need for human annotations. The recognition network, on the other hand, is an attentional encoder-decoder model that can predict text sequences directly from the rectified image.

2.3 Text Spotting

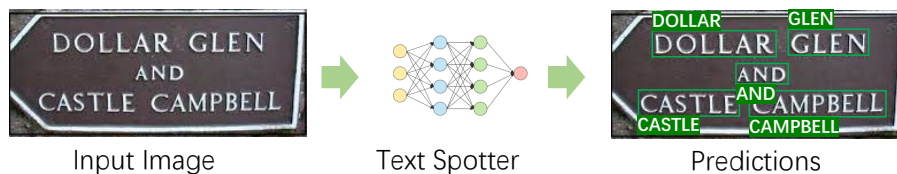
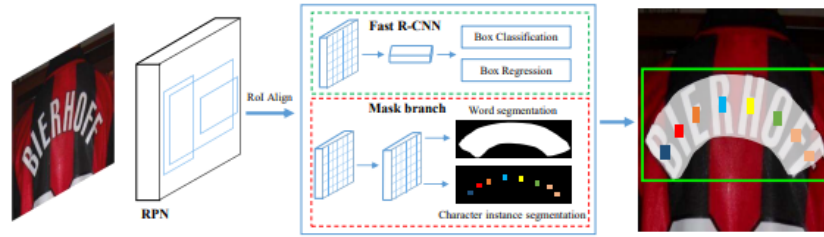


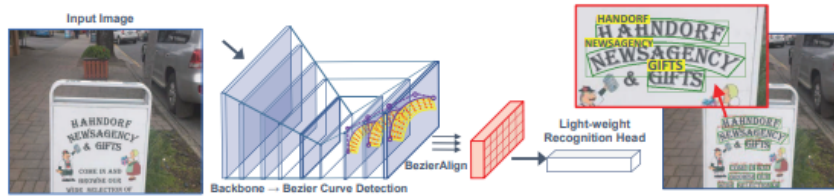
FIGURE 2.6. Text spotting, also known as end-to-end text detection and recognition, aims to simultaneously detect and recognize texts in images.

Considering the OCR task itself includes both text detection and text recognition as two sub-tasks. Simplifying and improving the system efficiency becomes a natural demand. To achieve this, end-to-end text detection and recognition, also known as text spotting (see Figure 2.6), was proposed. This is the task of simultaneously detecting and recognizing text in images within a single model, eliminating the separation of detection and recognition progress. Compared to traditional separated models, the text spotting model can improve system efficiency by reducing the number of computations and the need for post-processing steps. Additionally, it can also improve the overall accuracy of the OCR system by allowing the detection and recognition tasks to be trained together, enabling the model to learn the relationship between the two tasks. Therefore, it has attracted a lot of attention in recent years and has shown promising results in various applications [91, 102, 104, 109, 129, 133, 134, 172]. In general, there are two mainstream pipelines that are widely adopted for text spotting, i.e., segmentation-based methods and regression-based methods.

Segmentation-based text spotters treat text detection and recognition as a dense pixel prediction task. For example, Mask TextSpotter [109] (see Figure 2.7(a)) extends the well-known instance segmentation model Mask R-CNN [52] to a text spotter. Specifically, it generates character-level segmentation maps for each RoI and then utilizes a post-processing step to order these characters from left to right to group the final predictions. However, although it achieves



(a) MaskTextSpotterV3 [91] (segmentation-based)



(b) ABCNet [102] (regression-based)

FIGURE 2.7. Example methods of two mainstream pipelines of text spotters. (a) is a typical regression-based method ABCNet [102]. (b) is a popular segmentation-based framework MaskTextSpotter [91].

promising performance, the need for costly character-level annotation and time-consuming post-process steps makes it not well-suited for real-world applications where speed and efficiency are crucial. To tackle these issues, Liao et al. proposed an upgraded version of their previous work, called Mask TextSpotter v3 [91], which adopts a Segmentation Proposal Network (SPN) instead of the original RPN. Benefiting from the proposed SPN, Mask TextSpotter v3 enjoys a faster and more accurate performance than its predecessor while eliminating the requirement of character-level annotations.

Regression-based model is another widely-used approach for text spotting. It usually concatenates a recognizer to the end of a text detector and shares the same backbone features to improve efficiency. For example, an early regression-based text spotter proposed by Li et al. [85] replaced the object classification module in Faster-RCNN [141] with an encoder-decoder-based text recognition model. The recognition module can directly use the cropped backbone features, so the entire system can be trained in an end-to-end manner. Recently, Liu et al. [102] further improved the regression-based text spotter by introducing a more efficient ABCNet (see Figure 2.7(b)). The authors propose to use the Bezier curve to represent the text instances and design a novel Bezier Align module to share the backbone features between detection and recognition heads. Thanks to the concise representations and efficient feature-sharing

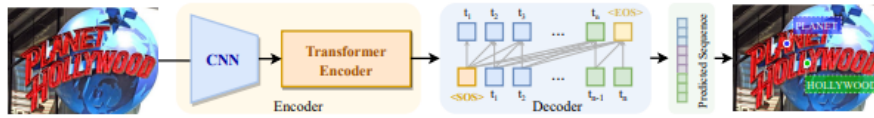


FIGURE 2.8. SPTS [129]

mechanism, ABCNet achieved promising performance with relatively low computational cost.

Although both segmentation-based and regression-based methods can be trained in an end-to-end manner and can simultaneously detect and recognize the text in images, they can still be considered two-stage methods. This is because both methods have a clear separation between the detection and recognition stages, even though they are combined into a single model. Within such a two-stage design, potential issues may arise, such as the need for a feature alignment module and a lack of interaction between the detection and recognition heads during training. More recently, Peng et al. [129] treated the text spotting as a sequence prediction task and proposed a novel text spotter called SPTS (see Figure 2.8). For the first time, SPTS merges the text coordinates with the text transcriptions together into a series of sequences. Then a transformer is employed to predict the sequence in an auto-regressive manner. Additionally, it also proves that the text instances can be represented by a single point rather than bounding boxes such as polygons, significantly saving the annotation costs.

In conclusion, while traditional two-stage text spotters such as segmentation-based and regression-based methods have shown good performance in end-to-end text detection and recognition, they still have limitations in terms of efficiency and interaction between the detection and recognition stages. Recently proposed ‘real’ end-to-end methods, such as SPTS, have shown promising results by eliminating the need for a feature alignment module and simplifying the representation of text to a single point. These methods with simple yet effective designs show potential directions for future research.

2.4 Downstream OCR Applications

The fact that OCR techniques can extract textual contents in images enables a variety of downstream applications, such as document analysis, handwritten mathematical expression recognition, and text visual question answering. These applications leverage the ability of OCR to accurately extract text from images, allowing for further processing and analysis. For example, document analysis can use OCR to extract structured information such as dates, names,

and numbers, which can be used for tasks such as automated data entry or searching through large collections of documents. Handwritten mathematical expression recognition can use OCR to convert written math equations into machine-readable formats, such as LaTeX sequence, enabling their use in digital documents. Text visual question answering can use OCR to extract text from images in order to answer questions about the text, such as "What is the title of the book in this image" which can be helpful for those visually impaired people.

Chapter 3

Benchmarking OCR systems: Datasets, Metrics, Methods

Statement of Authorship

Title of Paper	Benchmarking OCR systems: Datasets, Metrics, Methods
Publication Status	<input type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input checked="" type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	Submitted to Pattern Recognition

Principal Author

Name of Principal Author (Candidate)	Xinyu Wang		
Contribution to the Paper	Proposed the ideas, conducted experiments and draft the manuscript of the paper.		
Overall percentage (%)	90%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	19/Jan/2023

Co-Author Contributio

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

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

Name of Co-Author	Yuliang Liu		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	Jan/18/2023

Name of Co-Author	Chunhua Shen		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	Jan/18/2023

3.1 Introduction

Optical Character Recognition (OCR) is a long-standing research topic that has attracted tremendous interest from academia and industry. The fact that deep-learning-based methods can recognize text in various scenarios, especially those captured in challenging environments, has been an incredible development. However, an inevitable issue that hinders further improvements in this field is the lack of a standard evaluation protocol that allows fair comparisons from model to model.

Comparing the newly proposed methods with the state of the arts has been a litmus test for model effectiveness. For example, MS-COCO [93] serves as a touchstone for object detection, while most latest detectors report the corresponding performance. Most importantly, the experimental settings, such as backbone networks, schedules, and pipelines, are controlled to be as similar as possible to ensure fair comparisons. Such a fair comparison mechanism can be established not only because MS-COCO itself provides a robust evaluation protocol but also because of the accessible and reproducible configurations provided by the open-source community. Specifically, thanks to the standard detection libraries such as Detectron2 [179] and MMDet [19], it becomes pretty easy for researchers to develop and compare their models with previous methods by using consistent experimental settings. Unfortunately, a benchmark that can offer fair comparisons between OCR models is still inaccessible. Such a situation burdens researchers to expend extra effort to ensure consistent settings with others, which further impedes fair comparisons.

Figure 3.1 compares the typical evaluation protocols used by object detection and OCR tasks.

Diversity: The development of OCR is a bit diverging across various fine-grained scenarios, (*e.g.*, natural scene text [38, 184], handwritten text [78, 79], document text [63, 201], and digital picture text [53, 69]), all with their own datasets, metrics, and methods. Such divergence hinders further generalization of the OCR methods, *i.e.*, most of the existing models are designed for specific scenarios and can fail to work when the domain shifts. However, the fact that human beings do not need to consider writing forms (*e.g.*, printed or handwritten) and scenarios (*e.g.*, natural or digital) while reading text encourages the development of more generic OCR models, which the existing benchmarks cannot fully facilitate.

Volume: Another underlying problem has been that the dataset volume is much smaller than other vision tasks. Compared to object detection datasets,

Year/Venue	Method	Training Data	Syn.	Real	Backbone	Batch	Iteration	Solver	Test Size	Report
2019/ICCV	Qin et al. [135]	SynT200k, IC15, CT*, MLT17, TT, Pvt.	200k	56k+1M	Inception-R	15	8m	SGD	(600,)	70.7%
2019/ICCV	CharNet† [180]	SynT800k, TT	800k	1k	Hourglass88	32	125k+50k	SGD	(, 2280)	66.6%
2019/CVPR	CRAFT† [11]	SynT800k, TT	800k	1k	VGG-16	32	50k+25k	ADAM	(, 1280)	66.6%
2020/ECCV	MTSv3† [91]	SynT800k, IC13, IC15, SCUT, TT	800k	4k	R50-FPN	8	250k+250k	SGD	(1000, 4000)	71.2%
2020/CVPR	ABCNet [102]	SynT150k, MLT17, TT	150k	11k	R50-FPN	8	260k+5k	SGD	(1000, 1824)	67.1%
2020/AAAI	Qiao et al. [133]	SynT800k, TT	800k	1k	R50-FPN	8	500k+13k	SGD	(, 1350)	69.7%
2021/AAAI	PGNet† [167]	SynT800k, ArT*, TT	800k	6k	R50-FPN	48	N/A	ADAM	(, 640)	63.1%
2021/AAAI	MANGO [134]	SynT950k, IC13, IC15, CT*, MLT19, TT	950k	22k	R50-FPN	16	500k+250k	SGD	(, 1600)	72.9%
2021/MM	TDI [205]	SynT800k, IC13, IC15, SCUT, TT	800k	4k	R50-FPN	1	2.4m+600k	SGD	(, 1600)	70.1%
2021/PAMI	PAN++ [172]	SynT800k, IC15, CT, MLT17, TT	800k	72k	R18-FPEM	16	150k	ADAM	(736,)	68.6%

TABLE 3.1. Accuracy on a typical scene text dataset Total-Text [24] reported by recent proposed text spotters. Almost all of these OCR methods used external data to train the model within different training settings; This means that the comparisons between the reported performance on such benchmarks may not reflect the actual effectiveness of the proposed OCR algorithms. Datasets: SynT (SynthText) [49], IC13 [70], IC15 [71], ArT [25], TT (Total-Text) [24], SCUT [203], MLT17 [123], MLT19 [124], CT (COCO-Text) [38], Pvt. represents private data. Star* means that only part of the data is used. Dagger† means that character-level supervision is included at the training stage.

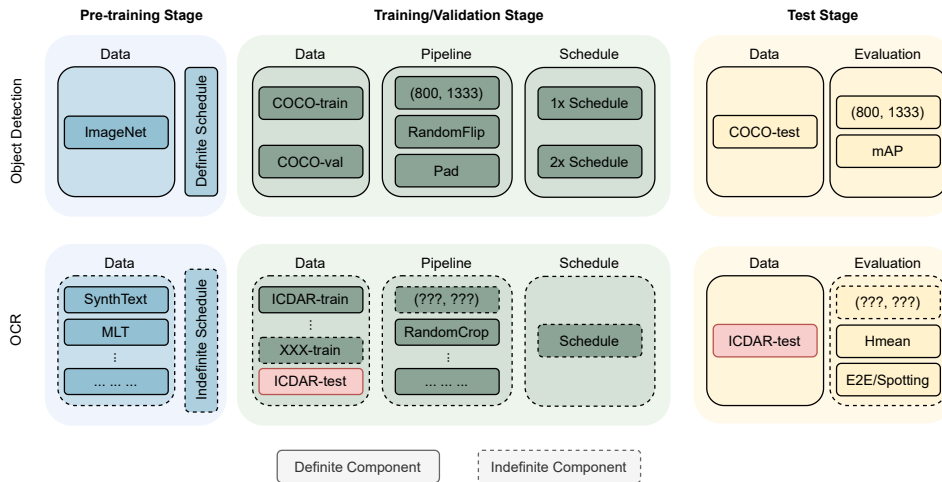


FIGURE 3.1. Comparison of typical evaluation protocols between object detection and OCR. The former usually uses definite experimental settings, including train/test datasets, data augmentation pipelines, training schedules, etc., to ensure fair comparisons, while the latter uses many indefinite configurations.

e.g., MS-COCO [93] with 118k/5k/41k images for train/val/test splits, three typical datasets that are used for scene text detection/recognition only include hundreds of samples, *i.e.*, MSRA-TD500 [184] with 300/200 for train/test; Total-Text [24] with 1200/300 for train/test; and SCUT-CTW1500 [100] with 1000/500 for train/test. Such a small volume of training sets enforces the OCR models to employ external data for training (see Table 3.1), and the use of different training data further leads to unfair comparisons. Besides, the absence of an official validation split and the insufficient testing images make it easy to cherry-pick the best-performed models on the testing set, inducing potential over-fitting. These issues increase the difficulty of a fair comparison; thus, the reported performance on such datasets may not reflect the actual effectiveness of the proposed models.

Annotation: Unlike the target in an object detection task that has a clear description and can be well-annotated by an axis-aligned bounding box or segmentation mask, the definition of instances in current OCR datasets can vary from case to case. From the perspective of language, most of the OCR datasets only label Latin characters, and the text presented in other languages is discarded as background or ignored region. From the perspective of labeling granularity, some datasets use word-level [24, 71] annotations, while others use line-level [28, 100, 184] or character-level [191] annotations. From the perspective of annotation form, rectangle [63, 115] and quadrilaterals [124, 197] are the most commonly used bounding box, while some other datasets utilize polygon

box [24, 100] to annotate the arbitrarily-shaped text. These labeling inconsistencies can be found among different datasets or inside a single dataset, which increases the cost and difficulty of cross-dataset training and evaluation.

In this chapter, we try to solve the above issues by integrating existing datasets and proposing a new evaluation system in order to fairly benchmark different OCR models. The contributions are three-fold:

- We systematically analyze the issues existing in the current OCR evaluation framework by conducting a series of ablation studies, identifying that unfair comparisons can significantly impact performance.
- Twenty-five publicly available datasets involving multiple scenarios and languages are collected to build a testbed, namely UniOCR, for benchmarking OCR methods. A benchmark suite is developed to provide fair comparisons between different OCR systems.
- Based on the UniOCR, state-of-the-art OCR algorithms are trained and tested within identical settings to build baselines. Analysis of experimental results is provided to explore insights from previous methods.

3.1.1 Related Work

Datasets: ICDAR 2003 [107] (IC03 for short) is one of the first well-organized OCR datasets, containing around 500 annotated images captured from natural scenes. Subsequently, a series of datasets that focus on scene text recognition was developed in the past decades, such as ICDAR 2011 [69], ICDAR 2013 [70], and ICDAR 2015 [71] (IC11, IC13, and IC15 for short). Later, Nayef et al. introduce one of the largest multi-lingual scene text datasets MLT-2017 [123], involving nine different languages; meanwhile, some non-English datasets such as Chinese [147], Indic [115], and Vietnamese [127], have also been proposed. Recently, recognizing arbitrarily-shaped scene text has been attracting more and more research interests; Total-Text [24] and SCUT-CTW1500 [100] are two popular benchmarks that contain text instances in curved shapes. In addition to natural scene text, OCR has also been adopted in many other scenarios, including documents [63, 201], handwritten text [78, 79], born-digital contents [53, 69], historical books [131], *etc.* For example, SROIE [56] collects one thousand receipt images to facilitate OCR and information extraction challenges for scanned receipts; FUNSD [63] introduces a dataset that contains hundreds of scanned forms; BID [29] proposes a large dataset for recognizing text in identity documents. Beyond static images, some datasets, *e.g.*, DOST [57] and LectureV-ideoDB [32], aim to recognize text from video clips, which further challenges

the OCR models to explore temporal information. Besides, there are also some datasets that were designed for detection [142, 184] only or recognition [40, 94] only tasks.

Metrics: There are mainly two steps to evaluate OCR systems; the first is to measure the localization accuracy, while the other is to assess recognition precision. For detection performance, supposing B_P and B_{GT} represents predicted and ground-truth bounding boxes, respectively. IC13 [70] employed a rule termed DetEval. Only when B_P and B_{GT} simultaneously satisfy $\frac{B_P \cap B_{GT}}{B_P} > \tau_1$ and $\frac{B_P \cap B_{GT}}{B_{GT}} > \tau_2$, the prediction would be considered as a true positive. IC15 [71] follows the same protocol used in the object detection benchmark Pascal VOC [34], which calculates the intersection-over-union value between B_P and B_{GT} . Only if the IoU value $\frac{B_P \cap B_{GT}}{B_P \cup B_{GT}}$ is larger than a designated threshold τ , the prediction and ground-truth are matched. Furthermore, some efforts [9, 99, 135, 147] have been made to diagnose and refine the existing metrics in the past years. For example, Liu et al. [99] propose TIoU to measure the tightness between predicted and ground-truth bounding boxes. Baek et al. [11] examine the inconsistencies of training and evaluation datasets for scene text recognition tasks. Once the predicted and ground-truth bounding boxes are matched, the textual content is compared to get recognition accuracy. Usually, there are two rules to calculate the recognition score. One compares whether the predicted text is exactly the same as the ground truth; the other uses a normalized edit distance score.

Methods: Modern OCR methods can be categorized into two groups, *i.e.*, end-to-end trainable text spotters [91, 102], and two-stage approaches [103, 146, 204]. The former detects and recognizes text instances in a unified network; the latter separates the OCR into two sub-tasks, *i.e.*, text detection and text recognition. End-to-end methods usually follow a multi-task joint training fashion. For example, Mask TextSpotter [91, 109] introduces character-level supervision to Mask R-CNN [52] which enables it to detect and recognize text and characters simultaneously. ABCNet [102] forms the irregular text instances with parameterized Bezier curves; the backbone features are shared by detection and recognition heads through a BezierAlign module, which significantly improves the network efficiency. For two-stage methods, text detectors [160, 204] usually adapted generic object detection frameworks to the text scenario, while text recognizers [146, 154] often stack RNN upon CNN to capture sequential features and trained with Connectionist Temporal Classification (CTC) loss. Some recently proposed recognizers [22, 108] employ attention mechanisms to extract two-dimensional features, which achieved impressive performance with

irregular text. Due to space constraints, we refer readers to [20, 106] for a comprehensive overview.

3.2 Unfair Comparison Between OCR Methods

Using similar training and inference settings is a prerequisite to ensure fair comparisons between different methods. However, after surveying a number of OCR algorithms reported on a widely used OCR dataset TotalText [24], we found that almost all of these models used external data and different training settings (see Table 3.1). Therefore, we cannot help asking how genuine that level of performance can the existing evaluation framework reflect. In this section, we employ the open-sourced OCR model ABCNet¹ [102] as a baseline to disclose the impact of unfair comparisons from the aspect of dataset, metric, and method.

3.2.1 Dataset Issues

Training: Due to the limited number of training samples in each single dataset, pre-training on synthetic samples [49] and external datasets has become indispensable step in training OCR models. Nonetheless, the lack of a unified standard caused chaos – dozens of combinations of external datasets can be found in different methods. For example, as shown in Table 3.1, an earlier OCR model introduced by Qin et al. [135] reported an impressive accuracy of 70.7% on the TotalText dataset. However, the model was trained on an extra 30k manually labeled private data as well as 1 million partially annotated images for 8 million iterations. In contrast, TextPerceptron, proposed by Qiao et al. [133], only uses synthetic data and 1k real samples for training but achieved a competitive accuracy of 69.7%. Comparing two models trained on completely different amounts of images is unfair and inappropriate; thus, the reported accuracy may not be convincing enough to reflect the actual performance. To explore the impact of the volume of training data, we conducted an ablation study. Table 3.2 shows that after excluding a part of the training samples, the accuracy of the baseline has declined to vary degrees. Specifically, the model that only used TotalText training split encountered a considerable performance drop in the End-to-End text spotting task, from an H-mean score of 67.1% to 51.9%, compared to the fully trained model.

Inference: The limited number of test samples (*e.g.* 300 images in TotalText [24]) makes the OCR models prone to overfitting, and the performance is,

¹<https://github.com/aim-uofa/AdelaiDet>

Training Set	Det-only (%)			End-to-End (%)		
	P	R	H	P	R	H
SynT, MLT, TT (Official)	91.5	80.0	85.3	69.9	64.5	67.1
SynT, TT	86.8	78.0	82.2	63.3	59.8	61.5
MLT, TT	82.7	78.9	80.8	53.2	52.9	53.1
TT	84.6	76.8	80.5	52.7	51.2	51.9

TABLE 3.2. Performance of ABCNet [102] on Total-Text [24] using different training data. Datasets: SynT [49], MLT [124], and TT [24].

Split	Det-only (%)			End-to-End (%)		
	P	R	H	P	R	H
Official	91.5	80.0	85.3	69.9	64.5	67.1
1	88.5	80.0	84.0	62.6	58.5	60.5
2	92.1	85.9	88.9	73.1	71.3	72.2
3	90.4	80.5	85.1	62.2	58.6	60.4
4	82.5	79.5	80.1	55.6	56.1	55.9

TABLE 3.3. Cross validation of ABCNet [102] on Total-Text [24].

therefore, susceptible to be affected once the test data moves beyond the original distribution. To quantitatively reveal the instability of accuracy evaluated on such datasets, we conducted cross-validation on TotalText [24]. Specifically, the entire dataset was separated into five splits, then each model was trained on four of five and tested on the rest split. As shown in Table 3.3, a significant gap of H-mean scores among different splits can be found, identifying that the accuracy on small test sets is not stable and thus might be arduous to reflect the actual performance in more generalized scenes.

3.2.2 Metric Issues

As shown in Table 3.4, most existing OCR datasets adopt the Pascal VOC [34] metric to match ground-truth and predicted bounding boxes, then using the full-match rule to calculate recognition score. However, such types of metrics can overestimate the detection performance while underestimating the recognition accuracy; hence, a large gap between detection and end-to-end accuracy can be observed, *e.g.* ABCNet achieves H-mean scores of 85.3% and 67.1% on detection and e2e task, respectively (see Table 3.2). We show two examples in Figure 3.2 and Figure 3.3 to explain this phenomenon. Specifically, the VOC metric measures the quality of predicted bounding boxes with a single fixed threshold (usually set to 0.5); therefore, the tightness between bounding boxes is ignored. For example, as shown in Figure 3.2, albeit the polygon prediction fits the GT bounding box better, all of the three predictions meet the IoU threshold

Dataset	Det. Metric	Rec. Metric
IC13 [70]	DetEval	Full-Match (Word-spotting)
IC15 [71]	VOC	Full-Match (Word-spotting)
MLT [124]	VOC	Full-Match (End-to-end)
Total-Text [24]	VOC	Full-Match (Word-spotting)
SCUT-CTW1500 [100]	VOC	Full-Match (End-to-end)

TABLE 3.4. Evaluation metrics used in some OCR datasets.

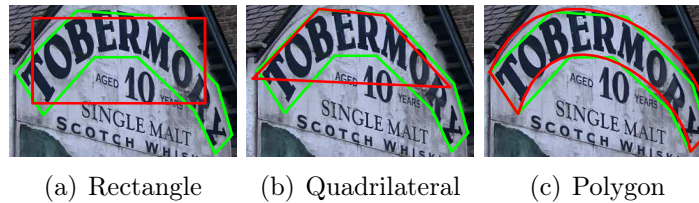


FIGURE 3.2. All of these predictions are treated as the same true positive under the VOC metric; thus, the detection performance is usually overestimated. Green and red bounding boxes represent GT and prediction, respectively.

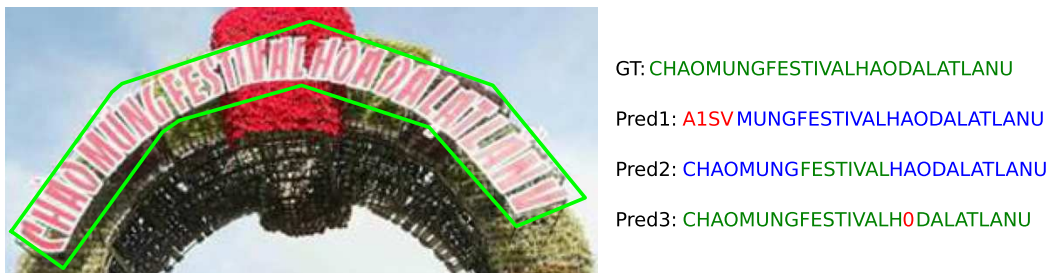


FIGURE 3.3. In most OCR benchmarks, the prediction contributes to the accuracy only if it is exactly the same as the ground truth, which sometimes underestimates the performance of recognizers.

and thus are considered equivalent in the evaluation process. Consequently, this metric can only get a relatively rough accuracy, which is unfair to the OCR algorithms that can obtain better-bounding boxes. When moving to the recognition side, the criteria become tougher. As shown in Figure 3.3, the accuracies of three predictions with different degrees of error will all be marked as zero, though Pred3 only misclassified one character while Pred1 totally failed to recognize the text. In this situation, the recognition performance cannot be differentiated, and the overall precision, especially for the longer instances, will be underestimated.

It is known that ordinary model settings such as the training schedule, image size, and batch size can play important roles in both the training and testing stages. Here, we use two examples, *i.e.*, input size, and recognition resolution,

Test Size	Det-only(%)			End-to-End(%)			FPS
	P	R	H	P	R	H	
Official	91.5	80.0	85.3	69.9	64.5	67.1	13.2
(600, 1200)	92.5	72.6	81.4	63.4	53.5	58.1	22.1
(800, 1600)	91.5	78.0	84.2	66.6	60.5	63.4	17.0
(1000, 2000)	91.4	80.0	85.3	69.7	64.4	67.0	12.9
(2000, 4000)	88.4	79.8	83.9	67.5	64.1	65.7	4.5

TABLE 3.5. Performance of ABCNet [102] on Total-Text [24] using different testing size. Official code uses (1000, 1824).

Rec. Resolution	CTW-E2E(%)			TT-E2E(%)		
	P	R	H	P	R	H
Official	57.4	48.4	52.6	69.9	64.5	67.1
(4, 32)	51.5	43.4	47.1	62.6	60.0	61.2
(8, 32)	53.4	44.7	48.7	-	-	-
(8, 128)	-	-	-	64.5	60.3	62.3
(16, 32)	49.9	41.3	45.2	65.6	63.7	64.6
(16, 128)	56.3	47.2	51.4	63.3	61.0	62.1

TABLE 3.6. Performance of ABCNet [102] on Total-Text [24] and SCUT-CTW1500 [100] using different recognition resolution. Official code uses (8, 128) and (8, 32) for CTW and TT, respectively.

to illustrate that such hyper-parameters can significantly impact the final performance. The commonly used solution to get input size at the inference stage is resizing the shorter or longer edge to a fixed length while preserving the aspect ratio, *e.g.* (, 1600) represents reshaping the longer side to 1600 pixels. Some methods may also set a maximum size to prevent inputting images that are too large; for example, (1000, 1824) denotes resizing the shorter edge to 1000 and keeps the longer side no larger than 1824. As shown in Table 3.1, the range of test image size can be significantly different from paper to paper, introducing obvious unfair comparisons. For instance, MTSv3 [91] sets a large maximum input size (1000, 4000), while PGNet [167] reshapes the images to only (, 640). To understand the possible impact caused by image size, we show the performance of ABCNet [102] within different input sizes in Table 3.5. When inputting smaller images with sizes (600, 1200), the end-to-end performance drops from 67.1% to 58.1%, and if a larger size (2000, 4000) is used, the model runs much slower. Another serious issue is the dataset-specific parameters which may be limiting the generalization ability. For example, Table 3.6 shows that the performance of the baseline model is sensitive to the data, as different hyper-parameters were respectively used to achieve the best performance on the two datasets, which, however, might not be the best settings for

ID	Year	Dataset	#Im	Granu.	Form	Lang.	Scenario
01	2011	IC11 [69]	0.5k	W	Rect.	EN	Digit
02	2018	MTWI [53]	20k	W	Quad	EN, CH	Digit
03	2013	IC13 [70]	0.5k	W	Rect.	EN	Nature
04	2015	IC15 [71]	1.5k	W	Quad.	EN	Nature
05	2021	TextOCR [152]	28k	W	Poly.	EN	Nature
06	2017	MLT [123]	10k	W	Quad.	Multi	Nature
07	2017	COCO-T [38]	63k	W	Poly.	EN	Nature
08	2017	R-CTW [147]	12k	L	Quad.	EN, CH	Nature
09	2019	ReCTS [197]	25k	W,L	Quad.	EN, CH	Nature
10	2019	ArT [25]	10k	W,L	Poly.	EN, CH	Nature
11	2019	LSVT [156]	50k	L	Poly.	EN, CH	Nature
12	2010	KAIST [81]	3k	C,W	Rect.	EN, KR	Nature
13	2017	ILST [115]	1k	W	Rect.	EN, HI	Nature
14	2021	VinText [127]	2k	W	Poly.	EN, VI	Nature
15	2016	DOST [57]	30k	W	Quad.	EN, JP	Nature
16	2019	FUNSD [63]	0.2k	W	Rect.	EN	Doc.
17	2019	SROIE [56]	1k	W	Quad.	EN	Doc.
18	2019	NAF [28]	0.8k	L	Quad.	EN	Doc.
19	2020	BID [29]	28k	L	Poly.	Latin	Doc.
20	2020	DDI [201]	100k	C,W	Quad.	Multi	Doc.
21	2015	DeText [189]	0.5k	W	Quad.	EN	Doc.
22	2021	GNHK [79]	0.7k	W	Quad.	EN	Handwrit.
23	2021	IMGUR [78]	8k	W	Rect.	EN	Handwrit.
24	2018	LV [32]	116k	W	Quad.	EN	Handwrit.
25	2016	HKWS [131]	5.5k	W,L	Rect.	Latin	Handwrit.

TABLE 3.7. UniOCR covers 25 publicly available OCR datasets. Granu. represents annotation granularity, including Char (C), Word (W), and Line (L). Licenses are included in supplementary.

other datasets. In summary, the above experiments demonstrate that hyper-parameters can heavily impact model performance; thus, it might be unfair to compare the methods using different settings.

3.3 Unified OCR Benchmark

The lack of a unified benchmark has led to considerable differences in the settings used in OCR methods, making it unclear whether the performance improvement is gained by algorithms or engineering tricks. Therefore, in this section, we will introduce a unified OCR benchmark. To our knowledge, this is the first trial to build a uniform and fair benchmarking protocol in the OCR community.

3.3.1 Data Collection

The aforementioned dataset issues have demonstrated that potential unfair comparisons can be introduced by training and testing samples. Therefore, the first step of building a unified OCR benchmarking system is to determine the image sets used for training, validating, and testing. To this end, twenty-five widely-used, publicly available, and well-annotated OCR datasets that contain both detection and recognition labels are collected from the existing literature (see

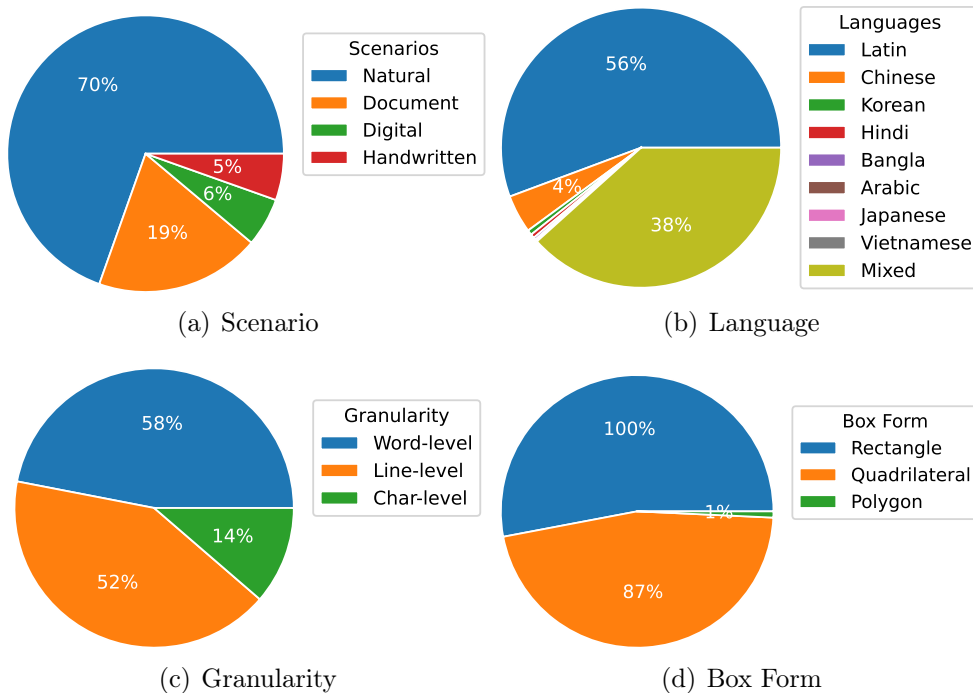


FIGURE 3.4. Data distribution of the UniOCR benchmark from the aspect of (a) Scenario; (b) Language; (c) Annotation Granularity; and (d) Bounding Box Form. It should be noted that as one image may contain multiple annotation granularity or bounding box forms, *e.g.*, Latin instances are labeled in word-level granularity while Chinese instances followed line-level labeling rules, the summations of percent values in (c) and (d) are larger than 1.

	Train	Val	Test	Total
Synthetic	272,972	-	-	272,972
Real	136,240	5,000	38,274	179,514
- Nature	92,014	3,531	25,676	121,221
- Digit	6,154	456	3,510	10,120
- Document	27,902	681	5,115	33,698
- Handwritten	10,170	332	3,973	14,475

TABLE 3.8. Data volume of UniOCR.

Table 3.7), we named this collection as UniOCR. To encourage the development of more generic OCR systems and evaluate their generalization ability, UniOCR covers images of four main scenarios, *i.e.*, natural scene, born-digital images, documents, and handwritten text, within text instances presented in more than ten languages, including Arabic, Chinese, English, Hindi, Korean, Russian, etc. Also, for each language, around 30k samples are synthesized using the SynthText [49] to facilitate possible pre-training of OCR models. It is noteworthy that some datasets do not release the annotations for test split

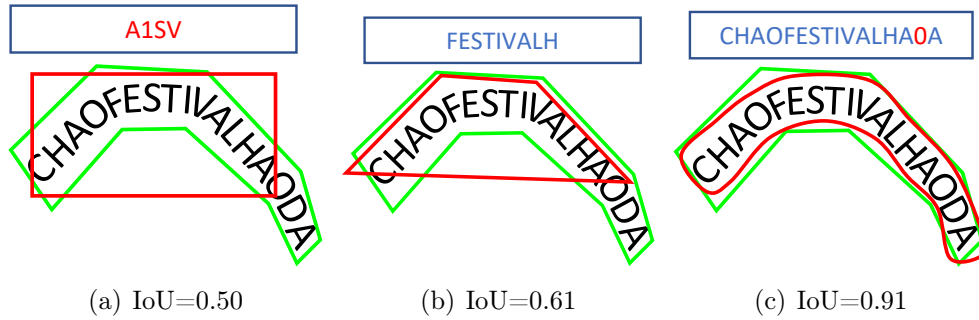


FIGURE 3.5. Detected boxes with different overlaps with GT.

Metric	Total-Text [24]		UniOCR	
	Box	Det	Det	E2E
(a)		1.0	0.1	0.0
(b)		1.0	0.3	0.1
(c)		1.0	0.9	0.9

TABLE 3.9. Precisions under Total-Text [24] and UniOCR metrics.

and/or do not provide an official validation split; in this situation, 20% images from the training set are separated into the val or test split. In addition, for some datasets with extraordinarily large sizes, such as LectureVideoDB [32], which offers more than 100k image frames extracted from video clips, UniOCR only sampled part of the data to ensure a reasonable data distribution. Finally, after filtering some invalid images with useless annotations, there are a total of around 180k images included in the UniOCR, which is split into the Train, Val, and Test sets (see Table 3.8). Figure 3.4 shows some detailed distribution of the dataset with respect to Scenario, Language, Annotation Granularity, and Bounding Box Form.

In order to ensure that identical training images and annotations could be accessed, unified interfaces and benchmark suites are developed, enabling automatic downloading, extracting, and processing of the data from the official project website of each dataset. Also, we will provide a pre-built version of the UniOCR, scripts used for converting to widely-used annotation formats such as COCO will also be included for research convenience².

3.3.2 Evaluation Metric

To alleviate the metric issues mentioned in Sec. 3.2.2, we adopt a combination of evaluation metrics in the UniOCR benchmark. For detection accuracy, we

²Code and pre-built version of UniOCR will be made public.

employ the COCO [93] Average Precision (AP) metric. Unlike the VOC metric used in most existing datasets (*e.g.*, Total-Text [24]) that only estimates the accuracy under a fixed threshold, AP considers the whole precision-recall curves; thus, the tightness between detection and GT bounding boxes can be measured to a certain extent. For example, Figure 3.5 shows three predictions with different IoU scores; however, they all reached the designated threshold ($\text{IoU} \geq 0.5$), and the precisions are thereby all 1.0. Under such metrics, the performance of detectors might be overrated (see Figure 3.5(a) and 3.5(b)). As shown in Table 3.9, when using AP metric, three detectors achieved 0.1, 0.3, and 0.9 AP (@0.5:0.95) scores, respectively, which well distinguishes the performance of different detectors according to the quality of predictions. For recognition accuracy, we employ an Averaged Normalized Edit Distance (ANED) score to assess the recognition accuracy for longer text lines. Specifically, the calculation of ANED is very similar to the detection AP, which replaces the IoU score with the normalized edit distance, and then precision under 10 thresholds of .50:.05:.95 are averaged, where the scores less than 0.5 will be directly set to 0. Table 3.9 shows that the recognition performance for longer text instances can be better evaluated under such metrics. The Harmonic mean (H-mean) between precision and recall is calculated for the final performance at the evaluation stage.

3.3.3 Preliminary Experiments

To understand the difficulty of UniOCR and provide guidance on the training settings for future methods, preliminary experiments based on ABCNet [102] are conducted. Models were trained on 4x Nvidia Tesla V100 GPUs with a total batch size of 8 within identical hyper-parameters as used in the official code.

3.3.3.1 Pre-training Matters

There are typically two ways for model pre-training, *i.e.*, synthetic-only and mixed training. The former only utilizes synthetically generated samples at the pre-training stage and then fine-tunes the model on real datasets, which is more commonly used by models trained with fewer real images (see Table 3.1). For example, TextPerceptron [133] was pre-trained on 800k synthetic images for five epochs, then fine-tuned on 1k real samples for the other 80 epochs. In the case that a larger size of training data is accessible, a mixed training pipeline becomes preferable [135, 172], *e.g.*, PAN++ [172] jointly trained their models on a combination of 800k synthetic + 72k real images, for 150k iterations without

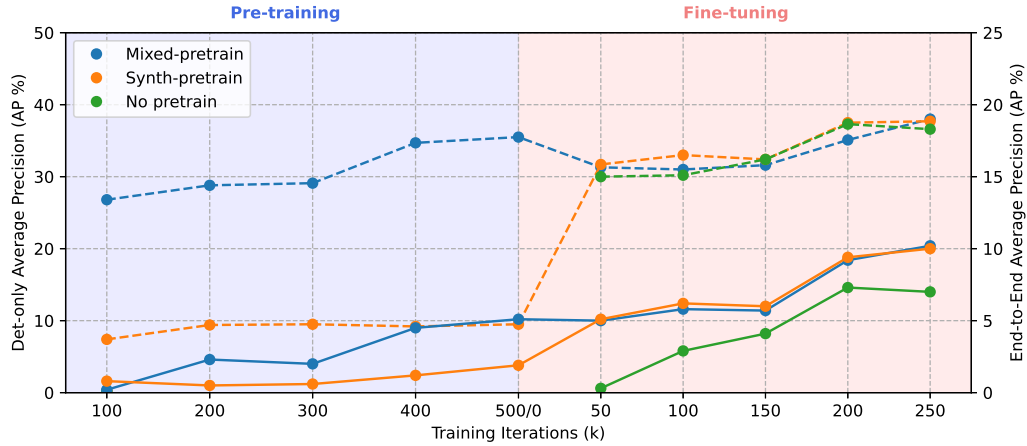


FIGURE 3.6. Performance of models using different pre-training strategies. Solid and dashed lines represent end-to-end and detection-only AP, respectively.

further fine-tuning. In addition, some other methods may use both strategies, for example, ABCNet [102] firstly pre-trained the models on a mixed dataset that involves both synthetic and real images for 260k iterations, then fine-tuned the model on real samples for other 5k iterations. Moreover, a recently proposed work [10] claims that synthetic samples are not necessary for recognition-only tasks when training with enough real images.

To explore the differences between the above pre-training strategies, we conducted ablation studies on the proposed UniOCR benchmark. Specifically, the mixed training model was pre-trained on a combination of synthetic real images from the UniOCR benchmark (see Table 3.8) for 500k iterations, then fine-tuned on real images only for other 250k iterations; while the synth-only model only used synthetic data at pre-training stage; for the real-only model, it was trained on real images without pre-training. The models were evaluated on the val split of UniOCR under the metrics introduced in Sec. 3.3.2. As shown in Figure 3.6, for the detection-only accuracy, three solutions, *i.e.*, mixed, synth-only, and no pre-train, have achieved quite similar performance, which is 38.0%, 37.7%, and 36.6%, respectively. It is noteworthy that the mixed-trained model has already achieved 35.5 AP without further fine-tuning, though 2.5% AP improvements were obtained after fine-tuning. Besides, the real-only model can obtain competitive results with only 50k iterations, suggesting that the text detection models can converge well on large-scale real datasets without external synthetic samples. When moving to the end-to-end text spotting task, both models that used synthetic data still achieved similar performance after fine-tuning (synth-only AP 10.0% *vs.* mixed AP 10.2%), however, outperforming the real-only model (AP 7.0%). Such results demonstrate that synthetically generated images are

Iterations	Document (%)		Handwritten (%)		Digit (%)	
	w/o Nat.	w/ Nat.	w/o Nat.	w/ Nat.	w/o Nat.	w/ Nat.
50k	42.5	30.2	44.0	31.8	41.7	36.1
100k	41.4	35.2	44.6	34.6	41.8	37.4
150k	45.5	37.3	43.9	35.0	40.8	41.4
200k	47.7	42.3	44.4	35.5	39.7	46.1
250k	47.7	42.8	43.8	42.8	39.4	46.6

TABLE 3.10. Detection AP of models trained with or without natural scene samples on subsets in different scenarios.

Iterations	Chinese (%)		Korean (%)		Vietnamese (%)	
	w/o EN	w/ EN	w/o EN	w/ EN	w/o EN	w/ EN
50k	38.0	35.5	32.3	15.1	46.6	38.0
100k	40.3	35.6	32.7	12.9	46.6	36.0
150k	43.0	37.3	33.1	21.8	47.0	38.2
200k	44.5	42.3	31.2	27.3	46.4	46.9
250k	45.0	43.1	30.8	25.7	46.4	47.3

TABLE 3.11. Detection AP of models trained with or without English text on subsets in different languages.

helpful for training end-to-end text spotters. Interestingly, the mixed-trained model achieved around 5.0% AP after 500k iterations, which is much worse than the real-only model that was trained for half iterations, suggesting that the end-to-end task takes more iterations to converge.

3.3.3.2 Scenario Matters

The fact that most existing OCR algorithms are only evaluated on single-scenario datasets limits the development of more generic methods. To explore the performance of current methods on more generic scenarios that contain text instances presented in different forms, we separately trained and tested the baseline method on four subsets of UniOCR, *i.e.*, natural scene, digital, handwritten, and document text. Table 3.10 compares the detection performance of models trained with or without natural scene text. Although the mixed scenario models use more training samples, they performed even worse, *e.g.*, the doc-only model achieved an AP of 47.7%, while the doc+nat model achieved an AP of 42.8%. Moreover, almost all of these single scenario-trained models obtained a better detection precision than the best model that was trained on the full dataset (see Figure 3.6). Such experiments suggest that those methods developed for a single scenario might be prone to learning domain-specific features, thus confusing handling data with a large domain gap. An exception is that the model mixed-trained with natural scene text images achieved better accuracy on the digital scenario, this may be because most of those digital pictures are obtained by rendering text instances to product and natural images

artificially; therefore, the mixed-trained model enjoys a performance improvement by involving the natural scene text subset. In summary, UniOCR enlarges its diversity by combining datasets from different scenarios, which challenges future methods to reduce the reliance on domain-specific knowledge and focus on more generic situations.

3.3.3.3 Language Matters

Similar to scenario matters, language can also be a factor to impact the OCR model performance. For example, unlike Latin words which are usually segmented by spaces, there is no apparent space between Chinese words. Therefore, Chinese text instances are often annotated with line granularity, while English text is with word granularity. Also, Semitic languages such as Arabic are written from right to left, which challenges the multi-lingual recognition performance. To this end, experiments are conducted to explore the language matters in UniOCR. As shown in Table 3.11, the mono-lingual models achieved slightly better accuracy on Chinese and Korean subsets while obtaining comparable precision on the Vietnamese subset. This may be because the Vietnamese text shares many similar characters to the English alphabet; however, the Chinese and Korean text is significantly different from Latin letters. Such experiments repetitively identified the generalization issues in current OCR models.

3.4 Experiments

3.4.1 Comparisons of State-of-the-art Methods

To further understand the difficulty and set a baseline for future methods, we have trained and tested some open-sourced state-of-the-art OCR algorithms [91, 92, 102, 146, 172] on the UniOCR. For a fair comparison, all models were trained with a batch size of 8 using the mixed-pre train strategy, *i.e.*, pre-trained on a combination of synthetic and real samples for 500k iterations, then fine-tuned for other 250k iterations on real images. Finally, all input images are resized to (1000, 1600) for evaluation at the inference stage.

Table 3.12 shows the H-mean score of averaged precision and recall of several open-sourced OCR methods. Based on the experimental results, some interesting findings can be observed.

- End-to-End methods achieve better E2E scores on English text instances but cannot outperform two-stage methods on Non-English samples; this

Methods	Val (%)			Test (%)																	
	Overall		Overall	Nature			Handwrit.			Digit			Doc.			English			Non-English		
	Det	E2E		Det	E2E	Det	Det	E2E	Det	E2E	Det	E2E	Det	E2E	Det	E2E	Det	E2E	Det	E2E	
MTSv3 [91]	22.2	11.0	22.9	11.3	19.6	8.3	31.0	18.3	23.7	3.4	23.6	14.3	21.1	13.0	25.5	9.3					
ABCNet [102]	41.6	23.3	41.8	23.2	38.2	19.1	38.5	23.9	49.7	12.8	44.5	27.9	38.6	26.4	46.6	18.9					
PAN++ [172]	20.3	9.7	21.2	10.8	19.6	11.0	22.9	13.9	29.9	16.8	20.3	7.5	21.9	12.0	20.5	9.4					
End-to-End Methods:																					
Two-stage Methods:																					
DB [92]+CRNN [146]	26.4	16.4	25.7	15.2	22.5	11.4	28.3	10.4	20.8	9.3	29.3	23.5	24.0	12.9	28.2	18.5					
DB [92]+SAR [87]	25.4	18.9	25.1	18.3	21.6	13.6	27.9	18.3	20.1	8.9	29.1	26.4	23.4	17.5	27.6	19.6					
PSENet [171]+CRNN [146]	20.4	13.0	19.8	11.9	15.3	8.7	23.5	10.2	17.9	10.6	24.0	16.3	21.8	12.0	17.8	11.8					
PSENet [171]+SAR [87]	20.2	15.9	19.8	15.2	15.2	10.2	23.6	16.9	17.6	9.8	23.7	21.6	21.8	17.5	17.6	12.7					

TABLE 3.12. Performance comparison between state-of-the-art OCR methods on UniOCR in terms of H-mean score.

Method	E2E (%)		
	P	R	H
CharNet (2019) [180]	-	-	66.6
PAN++ (2021) [172]	-	-	68.6
Text Perceptron (2020) [133]	-	-	69.7
TDI (2021) [205]	-	-	70.1
Qin et al. (2019) [135]	-	-	70.7
MaskTextSpotterv3 (2020) [91]	-	-	71.2
ABCNet (2020) [102] (official)	69.9	64.5	67.1
+ UniOCR pretraining	71.5	66.1	68.7
+ Large Input Size (1600, 3000)	72.6	67.0	69.7
+ Large Recog. Resolution (8x48)	70.9	69.3	70.1
+ Longer Schedule (2x)	75.1	69.2	72.0

TABLE 3.13. Performance on TotalText [24] without lexicon.

may be because the end-to-end approaches usually adopt a lighter recognition head that cannot discriminate the extensive dictionary well (there are around 8,000 unique characters included in the UniOCR).

- All methods achieved acceptable detection results on the Digit subset; however, failing on the recognition part catastrophically. The possible reason is that the Digit subset mainly consists of Chinese samples that are mostly annotated by text lines and have a large dictionary volume, which challenges the OCR models to generalize.
- Compared to other methods, ABCNet [102] achieves surprisingly high performance. This may be because the regression-based methods are less sensitive to the crowd-sourced labels, while the segmentation-based algorithms [91, 172] are easily affected by different annotation granularity. We show some visualized results in the supplementary material to qualitatively analyze such results.

3.4.2 Discussion

Why is UniOCR Necessary? Table 3.13 repetitively emphasizes the necessity of the proposed UniOCR benchmarking suite by showing that the model performance on the existing dataset can be easily manipulated. By adopting several training tricks, the ABCNet [102] gains significant improvements, outperforming other state-of-the-art algorithms. Such unfair tricks can be applied to any OCR model and therefore have less research value, making the reported precisions hard to reflect the actual performance. To this end, the proposed UniOCR excludes all irrelevant components by building a unified benchmarking

system, enabling fair comparisons between different methods, which genuinely unveil the effectiveness of candidate OCR models.

Limitation: A limitation that has not yet been solved in this chapter is the annotation granularity issue. Especially, UniOCR employs the original GT provided by each dataset to evaluate the model performance; hence, if a model predicts text lines for a specific image labeled at word level, the performance might be underestimated. To tackle this problem, new bounding box matching rules have to be carefully designed; we thereby leave this aspect for future work.

3.5 Conclusion

In this chapter, we have introduced UniOCR, an evaluation suite developed for benchmarking generic OCR algorithms. By combining twenty-five publicly available OCR datasets, UniOCR enjoys significant diversity in both scenarios and languages, thereby alleviating a number of issues existing in current evaluation protocols that hinder fair comparison between different methods. Preliminary experiments were conducted to explore insights and set a standard training schedule for the benchmark. Moreover, state-of-the-art methods involving both end-to-end text spotters and two-stage models were trained and tested on the UniOCR to set baseline results. To our knowledge, this chapter is the first attempt to introduce a unified benchmarking toolbox that enables fair comparisons in the OCR community, which we wish could become a valuable package for assisting future OCR-related research.

Chapter 4

Synthesizing High-Quality License Plates via a Text-to-Plate Network

Statement of Authorship

Title of Paper	Synthesizing High-Quality License Plates via a Text-to-Plate Network
Publication Status	<input type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input checked="" type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	Submitted to IEEE T-CSVT

Principal Author

Name of Principal Author (Candidate)	Xinyu Wang
Contribution to the Paper	Proposed the ideas, conducted experiments and draft the manuscript of the paper.
Overall percentage (%)	90%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	<hr/>
Date	19/Jan/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 to include the publication in the thesis; and
- iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Yuliang Liu
Contribution to the Paper	Discussion and writing revision.
Signature	<hr/>
Date	Jan/18/2023

Name of Co-Author	Chunhua Shen
Contribution to the Paper	Discussion and writing revision.
Signature	<hr/>
Date	Jan/18/2023

4.1 Introduction

License plate recognition is an essential task in intelligent transportation systems, attracting tremendous research interest from academia and industry. It has been widely used in many real-world applications, especially in traffic surveillance, self-driving automobiles, *etc.*

However, a long-standing issue in the license plate (LP) recognition task is the biases among the data collected from different sources, which induces the models trained on such datasets to a poor generalization ability. For example, CCPD [182] is one of the largest public datasets designed for LP recognition tasks, which contains nearly 300k images collected from 34 provinces in China. In the main track of the CCPD dataset, which is termed CCPD-2018 (see Fig. 4.1(a)), each image contains one blue Chinese LP (fuel vehicles), and the samples are further categorized into several splits, such as *base*, *tilt*, *weather*, and *rotate*. The updated version of this dataset, CCPD-Green (see Fig. 4.1(b)), further introduces a subset that contains around 10k green Chinese LPs (new energy vehicles). As shown in Fig. 4.1, both blue and green plates consist of a province code, a city code, and a series of ID numbers. The differences between these two types of plates are generally twofold: 1) the text colors are white or black while the background colors are blue or green, respectively; 2) the length of ID numbers are 5 and 6, respectively. However, such a small difference introduces a large domain gap between these subsets. In specific, a model that achieved state-of-the-art performance on a single subset can hardly be generalized into the other, which can only obtain almost 0 accuracies without fine-tuning (see Sec. 4.4 for detailed discussion). Meanwhile, the license plate data can also be sensitive to the region where the data is being collected. Fig. 4.2 shows that there are more than 95,000 photos collected from the same province, which means the majority of instances have an identical region code while the rest of the regions may only have 1-10 pieces of samples. Such a long-tailed distribution limits the generalization ability of the models trained on this dataset.

To solve the aforementioned issue, a common strategy is to synthesize a large number of images for pre-training the recognition model. The simplest and the most widely-used way to generate synthetic LP images is to randomly render characters and digits on a blank plate template [16]. However, the data that is generated by such methods remains a significant domain gap from the real dataset. Hence, the recognition models can only gain limited improvements from these generated images.

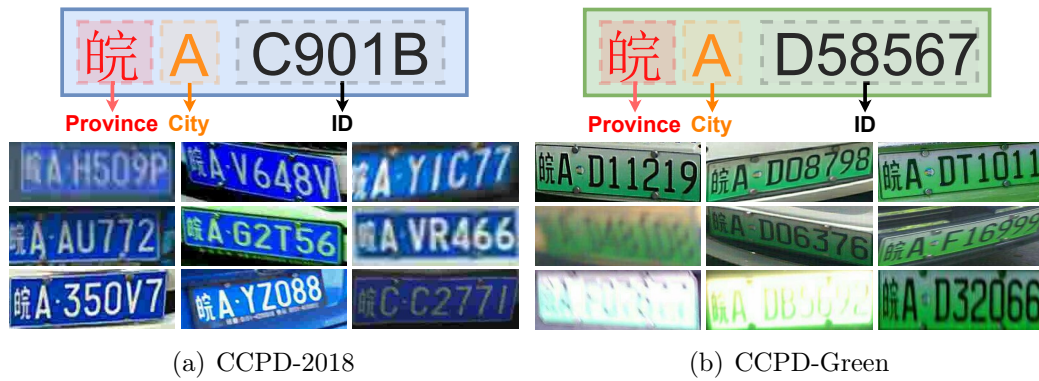
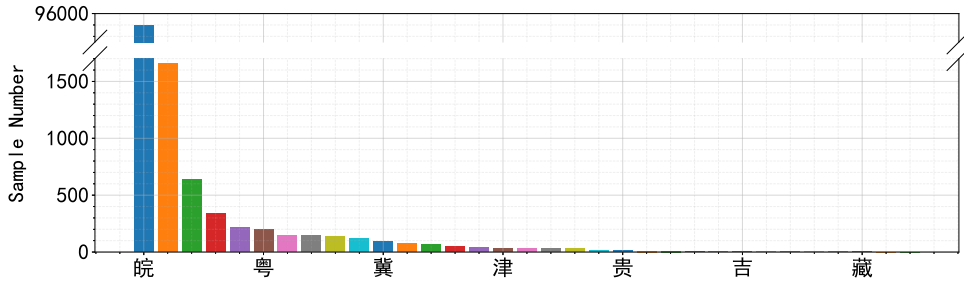


FIGURE 4.1. Examples of license plates from the CCPD dataset [182]. (a) CCPD-2018 contains 300k blue license plates, and each consists of 1 province code, 1 city code, and 5-char ID. (b) CCPD dataset contains 10k green license plates, and each consists of 1 province code, 1 city code, and 6-char ID.

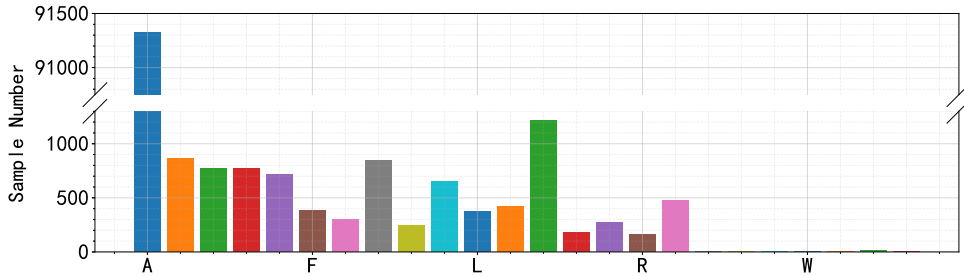
Some recent works [14, 196] try to enhance the LP synthesizing procedure. For example, [196] employs a CycleGAN [207] model to adapt the synthesized images to a photo-realistic domain. These methods have improved the quality of synthesized images to a certain extent, but they still cannot generate samples that can replace real data, since the models trained only using the samples generated by these methods failed to generalize to the real-world images. Therefore, we proposed a novel license plate synthesis method termed TLPNet based on a text-to-image framework. TLPNet can produce high-quality training samples, enabling the recognizer to achieve competitive performance to the model trained on real data. Especially for the long-tailed data, a baseline recognizer can only achieve 51.0% performance on the minority split on CCPD (excluding all LPs starting with “皖” in test split) when trained on 5k real images annotated by humans. However, the performance significantly raised to 80.2% when training the same model on 1m synthetic data generated by a TLPNet trained on a similar volume of real images without further fine-tuning on real images.

The main contributions of this chapter are as follows:

- We proposed a novel Text to License Plate Network (TLPNet) to enable the generation of high-quality synthetic LP images. The proposed TLPNet can generate significantly high-quality images compared to other text synthesis techniques.
- Based on the proposed TLPNet, we introduce the TLP-Syn dataset, which comprises 1 million synthetic LP images generated by the proposed TLPNet. The TLP-Syn dataset can be used as a supplement dataset for pre-training.



(a) Sample distribution based on province code.



(b) Sample distribution based on city code.

FIGURE 4.2. Data distribution of the CCPD-2018 [182] training split shows the majority ($\sim 90\%$) of car license plate samples are collected from a single city, which has identical province and city code (皖A).

- Comprehensive experiments show that TLPNet significantly outperforms various existing license plate synthetic methods by a large margin. Specifically, the performance of our method can be near twice that of CycleGAN using the same amount of synthetic data.
- To evaluate the effectiveness of the proposed modules, we thoroughly conducted experiments on several widely used benchmarks and achieved state-of-the-art performance.

4.2 Related Work

4.2.1 License Plate Recognition

License plate recognition can be regarded as a fine-grained version of OCR tasks with relatively fixed formats. Therefore, most license plate recognizers follow a similar pipeline to the generic recognition model designed for scene text recognition. These models can usually be further separated into two categories, segmentation-based methods [16, 39, 42, 88, 149] and non-segmentation-based methods [84, 86, 182, 196]. Segmentation-based methods usually predict character-level masks for the LPs, then feed these individual characters

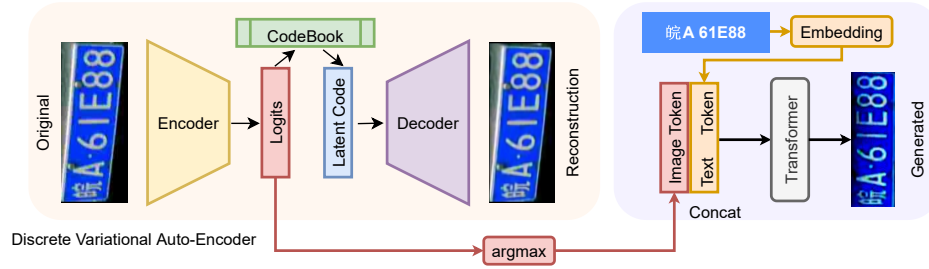


FIGURE 4.3. Structure of the Text to License Plate Network (TLPNet). The TLPNet consists of three parts: 1) a dVAE that compresses images into tokens; 2) an embedding layer that encodes each input license plate string into textual features; 3) a transformer that aggressively models the combined image and text features.

into a generic OCR model to obtain the recognition results. For example, [42] extracted character-specific extremal regions as character regions, then a hybrid discriminative restricted Boltzmann machine is employed for recognizing the characters. [16] proposed an end-to-end system for segmentation and an annotation-free license plate recognition system, which combined the segmentation and OCR modules by employing hidden Markov models. Then applying the Viterbi algorithm selects the most likely character sequence.

It is admitted that although the segmentation-based methods show superior performance on arbitrarily-shaped scene text, they usually have low inference speed and are susceptible to noises, including lighting, blurring, *etc.* However, as the license plate numbers are often well aligned, such approaches that work well on arbitrary shapes show fewer advantages for license plate recognition.

Therefore, the majority of recent works are developed based on a segmentation-free fashion. For example, [84] forms LP recognition as a sequence reading task, which proposed an LSTM-based sequence labeling network to recognize sequential features extracted by CNN. Then, the probability estimation sequence outputted by the LSTM is decoded by Connectionist Temporal Classification (CTC). [182] proposes a Roadside Parking net (RPnet) to detect and recognize LPs jointly. In the recognition part, deep features captured from different levels of convolutional layers are simply concatenated for training plate number classifiers. [86] encodes RPN extracted region features by bidirectional RNNs, then decodes the features into sequence by CTC, which allows a unified detection and recognition manner. [196] introduces a widely used attention mechanism into the LP recognition system, enabling the network to estimate the character location by 2D attention map rather than image appearance, further improving the robustness to arbitrarily-oriented instances.

4.2.2 Synthetic Text Generation

The surge of deep learning has significantly advanced scene text recognition performance in recent years. As collecting and annotating real samples can be highly expensive and time-consuming, pre-training the models on synthetically generated data has become a de facto standard pipeline for text recognition. Benefiting from the synthesis algorithms, it becomes easy for the models to access massive training data at an acceptable cost. Therefore, designing text synthesis methods has become an important research topic among the OCR community [2, 49, 59, 61]. [59] is one of the first extensive synthetic datasets designed for scene text recognition tasks, consisting of more than 9 million images covering 90k English words. Precisely, a candidate word from the dictionary is first rendered on a blank canvas within a randomly picked font, then a series of image processing operations such as prospective distortions, shadow rendering, coloring, and blurring is applied to improve the data diversity. Contrary to [59] that directly renders text on blank backgrounds, [49] overlays synthetic text to natural photos, accounting for the background 3D scene geometry. In specific, a natural image is first segmented into contiguous regions; local surface normal is then estimated based on predicted dense pixel-wise depth maps for each contiguous region; finally, a text sample is rendered to the local surface orientation within random fonts and transformations. This procedure enables the synthetic engine to produce more realistic scene text with natural backgrounds.

Synthetic data is also widely used in the LP recognition system [14, 16, 196]. The straightforward way is rendering the characters and numbers to a blank LP template directly [14, 16], accompanying morphology operations such as perspective transforming to add noises. The advantage of such types of methods is that they are free from manually annotated labels. However, it is noteworthy that there is a large appearance gap between the images synthesized by these methods and real photos. To alleviate this issue, a possible way is to use the existing labeled images to aid the generation of more realistic synthetic samples. For example, [196] proposes an AsymCycleGAN to adapt the generated LPs to a photo-realistic domain, improving synthetically generated images' quality. Our work also falls into this category. Nevertheless, we significantly improved the quality of the synthesized images by examining the rich features among the existing datasets using a transformer-based text-to-image framework, which enables a zero-shot text-to-image generation of more realistic LPs.

4.2.3 Text-to-image Generation

Translating descriptions formed in natural language directly into image pixels is a challenging task that lies at the intersection of computer vision and natural language processing, which has been attracting great interest from researchers in recent years [83, 114, 126, 137, 140, 181, 193, 194]. In general, the mainstream approaches of text-to-image generation can be found as two splits, *i.e.*, GAN-based [83, 140, 181, 193, 194] and Non-GAN-based [114, 126, 137] models. Specifically, [140] is a pioneering work that applied GAN [41] to generate plausible images from detailed text descriptions. Text embeddings are encoded by a hybrid character-level RNN, and then both the generator and discriminator networks perform feed-forward inference conditioned on the text features. [193] and [194] decompose the text-to-image generation into two sub-tasks following a coarse to fine fashion. A two-stage StackGAN is proposed to iteratively generate photo-realistic images, where the Stage-I network roughly generates low-resolution outlines; then, the Stage-II network learns to capture the details that are omitted by the former progress and draw high-resolution pictures. [114] is one of the first Non-GAN-based methods, which illustrated that the Deep Recurrent Attention Writer (DRAW) [47] can generate novel visual scenes when extended to the condition on image captions. [126] introduces an additional prior on the latent code, which significantly improves both the image quality and diversity of the synthesized samples. [137] developed a two-stage Vector Quantized-Variational Auto-encoder (VQ-VAE). In the first stage, a discrete variational auto-encoder (dVAE) is trained to learn the visual codebook, which compresses each RGB photo into a small image token. In the second stage, an autoregressive transformer is trained to learn the joint distribution of BPE-encoded text tokens and image tokens that were obtained from the first stage. Different from previous methods, which mainly pay attention to the natural scenes, our work is the first one that focuses on scene text-related generation.

4.3 Methods

4.3.1 Text to License Plate Network

Inspired by the recent success of text-to-image generation [137, 138] achieved by dVAE and transformer, we propose a Text to License Plate Network (TLPNet) to synthesize car license plates. As shown in Fig. 4.3, the TLPNet is composed of three parts: 1) a dVAE that compresses input images into tokens; 2)

an embedding layer that encodes LP strings into textual features; 3) a transformer that autoregressively models the joint distribution between image and text features.

The overall procedure can be considered as maximizing the variational lower bound [74]. Assuming that dataset X consists of observations while Z are unobserved hidden variables. Based on the Bayes' Theorem, the posterior distribution of Z can be obtained by $p(Z | X) = \frac{p(X|Z)p(Z)}{p(X)} = \frac{p(X|Z)p(Z)}{\int_Z p(X,Z)}$. Thus, the marginal probability of X can be written as $\log p(X) = \log \int_Z p(X, Z)$. As the true posterior density $p(Z | X)$ is intractable, it has to be approximated by a learnable distribution $q(Z)$. It is noted that the equation $\log p(X) = \log \int_Z p(X, Z) \frac{q(Z)}{q(Z)}$ still holds when adding $q(Z)$, and thus it can be regarded as an arbitrary distribution:

$$\log p(X) = \log \left(\mathbb{E}_q \left[\frac{p(X, Z)}{q(Z)} \right] \right). \quad (4.1)$$

Applying the Jensen's inequality $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$ on concave function $\log(\cdot)$, we can get:

$$\begin{aligned} \log p(X) &\geq \mathbb{E}_q \left[\log \frac{p(X, Z)}{q(Z)} \right] \\ &= \mathbb{E}_q[-\log q(Z) + \log p(X, Z)], \end{aligned} \quad (4.2)$$

which is termed the variational lower bound. Let us denote $\mathcal{L} = \mathbb{E}_q[-\log q(Z) + \log p(X, Z)]$, it is obvious that \mathcal{L} is a lower bound of the marginal probability of X . Therefore, instead of maximizing the marginal probability, it is convenient to maximize its variational lower bound \mathcal{L} .

Since the core idea of variational inference is to find the approximation distributions $q(Z)$ that are as close as possible to the true posterior likelihood $p(Z | X)$, it becomes necessary to measure the relative entropy between $q(Z)$ and $p(Z | X)$. To measure the differences between two probability distributions, the Kullback-Leibler (KL) divergence is one of the widely-used metrics. For distributions $q(Z)$ and $p(Z | X)$, the KL divergence is defined as:

$$\begin{aligned} \text{KL}(q(Z), p(Z | X)) &= \int_Z q(Z) \log \frac{q(Z)}{p(Z | X)} \\ &= - \int_Z q(Z) \log \frac{p(X, Z)}{q(Z)} + \log p(X) \int_Z q(Z). \end{aligned} \quad (4.3)$$

Combining Eq. 4.2 and Eq. 4.3, the variational lower bound can be finally rewritten as follows:

$$\mathcal{L} = \log p(X) - \text{KL}(q(Z), p(Z | X)). \quad (4.4)$$

TABLE 4.1. Network structure of the discrete variational auto-encoder.

Encoder		Decoder	
Conv (4 × 4, 256)		Conv(1 × 1, 256)	
Conv (4 × 4, 256)		ResBlock	Conv (3 × 3, 256)
ResBlock	Conv (3 × 3, 256)		Conv (3 × 3, 256)
	Conv (3 × 3, 256)		Conv (1 × 1, 256)
	Conv (1 × 1, 256)		Conv (3 × 3, 256)
	Conv (3 × 3, 256)		Conv (3 × 3, 256)
	Conv (3 × 3, 256)		Conv (1 × 1, 256)
	Conv (1 × 1, 256)		Conv(4 × 4, 256)
Conv (1 × 1, 8192)		Conv(4 × 4, 256)	Conv(1 × 1, 3)

Furthermore, in our case, we maximize the variational lower bound on a joint likelihood of the model distribution over images x , LP strings y , and the image tokens z encoded by the dVAE. Specifically, the distribution is modeled by the factorization:

$$p_{\theta,\psi}(x, y, z) = p_{\theta}(x | y, z)p_{\psi}(y, z), \tag{4.5}$$

where p_{θ} refers to the decoder in the dVAE, since it reconstructs a distribution over the possible corresponding values of RGB image x based on the encoded image tokens z . p_{ψ} denotes the joint distribution between the image and text features modeled by the transformer. Eq. 4.5 yields the variational lower bound:

$$\log p_{\theta,\psi}(x, y) \geq \mathbb{E}_{z \sim q_{\phi}(z|x)} (\log p_{\theta}(x | y, z) - \text{KL}(p_{\psi}(y, z), q_{\phi}(y, z | x))), \tag{4.6}$$

where q_{ϕ} is the encoder in the dVAE. Given an RGB image x , q_{ϕ} generates a distribution over the possible values of the latent representation z .

4.3.1.1 Variational Auto-encoder

Directly using image pixels as inputs for training the text-to-image generator would require an outrageous amount of memory. Thus, it is necessary to compress the raw images x into smaller tokens z . To this end, we train a discrete variational auto-encoder (dVAE) to transfer each input RGB image from $128 \times 128 \times 3$ into a 32×32 grid of tokens, which significantly reduces the size of features without noticeable degradation of image quality.

The dVAE consists of an encoder network $z = \phi(x)$ and a decoder network $\hat{x} = \theta(z)$. For each raw image x , $\phi(\cdot)$ encodes it into a latent representation vector z , while $\theta(\cdot)$ is responsible for reconstructing the encoded token z from latent space to the original space, \hat{x} is the reconstructed image. Combining these two parts together, the entire model of dVAE can be described as $\hat{x} = \theta(\phi(x))$. Specifically, both $\phi(\cdot)$ and $\theta(\cdot)$ are convolutional neural networks. As shown in Table 4.1, both are composed of a series of convolutional layers and residual blocks [51]. Primarily, the encoders and decoders use the convolutions with kernel size 3 and skip connections with kernel size 1, while others use 4×4 convolutions. The last layer of the encoder produces $32 \times 32 \times 8192$ output as the logits to represent the categorical distributions of the image tokens.

Given a raw image x , our goal is to obtain the latent code z . Supposing the posterior for the latent space is $p(z|x)$, the goal can then be expressed using the Bayes' Theorem $p(z|x) = \frac{p(x|z)p(z)}{p(x)}$. Based on Eq. 4.2, Eq. 4.3, and Eq. 4.4, the objective of the auto-encoder can be computed by minimizing the loss function:

$$L = -\mathbb{E}_{z \sim q(z|x)} (\log p(x|z) + \text{KL}(p(z), q(z|x))), \quad (4.7)$$

where the $q(z|x)$ and $p(x|z)$ are approximated by the encoder and decoder, respectively. The first term is the reconstruction loss, and the second can be viewed as a regularization term of the posterior.

Besides, as the transformer for generating images works on discrete data, it is necessary to convert the continuous latent representation learned by the standard VAEs to a discrete one. A common way to solve this issue is by adding a discrete codebook to the network. The codebook can be regarded as a lookup table that stores a list of vectors. Each output of the encoder $\phi(x)$ is compared to all the vectors in the codebook, then the codebook vector closest to the $\phi(x)$ in the euclidean distance is further passed to the decoder. The quantized vector can be calculated as follows:

$$\hat{\phi}(x) = \operatorname{argmin}_i \|\phi(x) - c_i\|_2, \quad (4.8)$$

where c_i is the i^{th} vector in the codebook. Practically, we use an embedding layer to maintain a codebook of size 512, and each element c_i in the codebook is a 32×32 vector. Thus, it can produce $512^{32 \times 32}$ different results, allowing the dVAE to learn a discrete space representing the raw image x well.

In practice, as the images of LPs are relatively simple, we directly use Mean Squared Error (MSE) as the objective function to optimize the dVAE reconstruction loss instead of combining a logit-Laplace distribution along with the

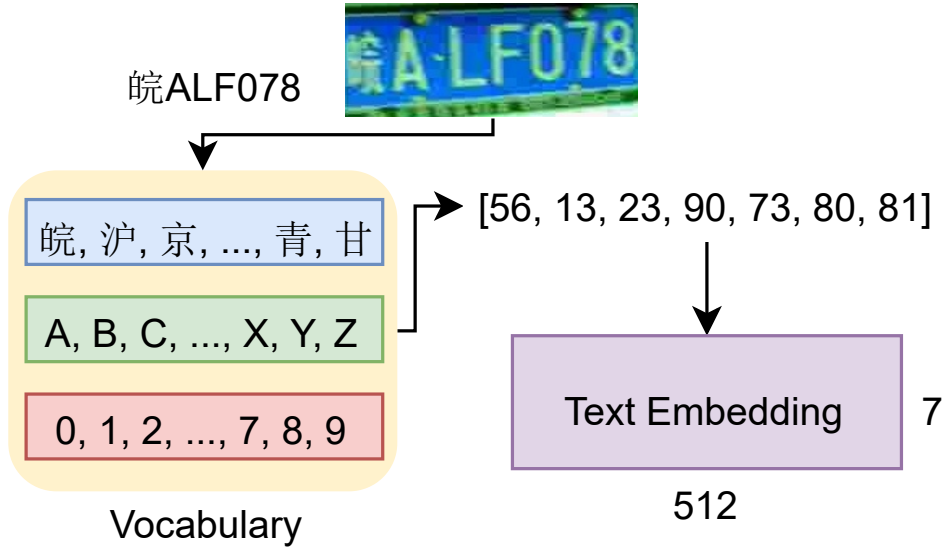


FIGURE 4.4. License plate number embedding.

KL divergence. As shown in Fig. 4.3, given an input image x_i , the dVAE encoder first extracts the convolutional features, then the logits are encoded into the latent code by the learned codebook. Finally, the decoder decodes the latent code into the reconstructed image \hat{x}_i . The final loss for training the dVAE can be relaxed to:

$$\ell = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}_i)^2. \quad (4.9)$$

The model is optimized by an Adam solver [73] with an initial learning rate of 0.001 for 20 epochs.

4.3.1.2 Transformer

After the training of dVAE, we can learn a new prior $\hat{p}(z)$ that can accurately describe the distribution of latent space, thus, when we sample data from $\hat{p}(z)$ and feed them to the decoder $\theta(\cdot)$, a new image can be generated $\hat{x} = \theta(\hat{p}(z))$. In this stage, we employ a sparse transformer [23] to learn the prior $p_\psi(y, z)$ distribution over the license plate text y and the image tokens z obtained by the dVAE. Since we have already obtained the encoder p_ϕ and decoder p_θ in the trained dVAE, based on Eq. 4.5 and Eq. 4.6, we only need to learn the prior by maximizing the variational lower bound with respect to ψ while fixing ϕ and θ .

Unlike other text-to-image generation tasks, the format of LPs is relatively simple, and does not carry rich language semantics; thus, it is not necessary to employ complicated language models to obtain the word embedding for the text descriptions. Alternatively, we use a simple yet effective way to convert each LP number to text embedding (see Fig. 4.4). Specifically, the LP number

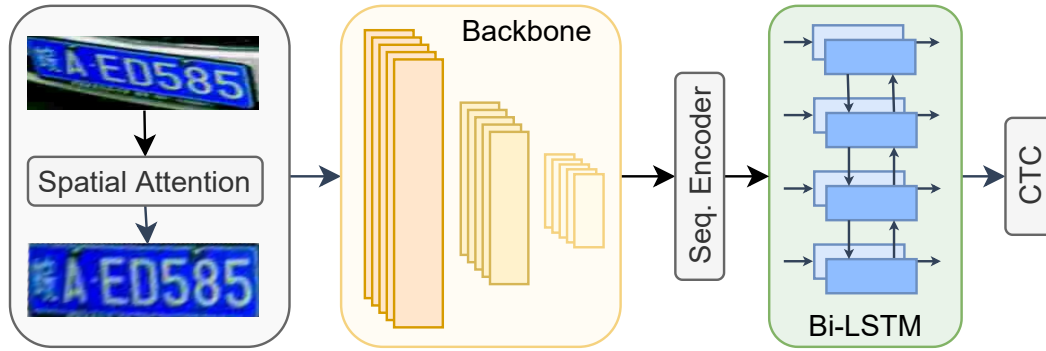


FIGURE 4.5. The network structure of the baseline recognizer, is composed of three: spatial transformer, backbone network, and bidirectional LSTM.

in the CCPD dataset is composed of three parts, *i.e.*, province code, city code, and ID. The province codes are of 34 Chinese characters, while city codes and ids are of English characters and digits. Each character is indexed with a unique integer in a dictionary, based on this, the LP number is transferred into a 7-bit digit vector. Furthermore, it is easy to maintain a simple lookup table that stores embeddings of a fixed dictionary and size. Therefore, given a list of indices, the lookup table can retrieve and output the corresponding word embeddings via indices. It is noted that the lookup table can be easily implemented with a linear layer and optimized as part of the training task. So far, given a pair of images and LP numbers, the image tokens can be easily sampled from the aforementioned dVAE encoder logits, and the plate number is converted into text embedding via the lookup table. Finally, these features of the two modalities are concatenated together as a single stream of feature tensors.

The combined feature tensors are fed into a decoder-only transformer to model the joint distribution between text content and image tokens. The transformer is composed of 32 self-attention layers, each of which uses 16 attention heads. The model uses a mixture of sparse attention masks, similar to [23, 137]. Specifically, the image receives three types of sparse attention, *i.e.*, axial attention along the rows and columns, respectively, as well as convolution-like attention. At the same time, the text embeddings are always obtained full attention. We refer interested readers to [137] for more details. During training, we use a reversible residual layer [75] instead of a regular one to scale the transformer depth, which significantly saves the memory cost, thus enabling a larger batch size of 24 images on a single V100 GPU. Specifically, the transformer was optimized using Adam with the cross-entropy loss for 75 epochs.

4.3.2 A Simple Yet Strong Recognizer

To alleviate the possible impact introduced by complicated recognizer while exploring the effectiveness of the proposed synthesizing methods, a simple and clean framework is developed as our baseline recognizer, which still achieves state-of-the-art performance on all 4 benchmarks without bells and whistles.

As shown in Fig. 4.5, there are mainly three modules in the baseline recognizer, *i.e.*, spatial attention mechanism, backbone network, and bidirectional LSTM (biLSTM). The spatial attention mechanism rectifies the loosely bounded LP region into a tightened one. The backbone network extracts deep convolutional features, which are then encoded into sequences. Finally, the sequential representation extracted by the biLSTM is decoded by Connectionist Temporal Classification (CTC) [44].

More specifically, the spatial transformer takes a grey image $\mathbf{I} \in \mathbb{R}^{W \times H}$ as input, then predicts an affine transformation matrix:

$$\theta(\mathbf{I}) = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}. \quad (4.10)$$

The matrix predictor can be simply implemented by several convolution layers and a regression layer. Similar to Spatial Transformation Networks (STN) [60], the regressor does not get direct supervision towards the parameters of affine transformation but will be trained according to the recognition loss. After obtaining the transformation matrix, each pixel (x'_i, y'_i) in the original image is sampled to the rectified space (see Eq. 4.11).

$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} = \theta(\mathbf{I}) \begin{bmatrix} x'_i \\ y'_i \\ 1 \end{bmatrix}. \quad (4.11)$$

After that, four pixels in the original image nearest to the sampled point (x_i, y_i) are used to generate the intensity values for the output image by bilinear interpolation. Furthermore, the rectified images are fed into a ResNet-34 [51] backbone network for extracting deep features.

As the text is inherently a sequence of characters, it is common to regard text recognition as a sequence reading task in the OCR community [96, 144]. Therefore, to explore the long-range dependencies among the sequences, we employ bidirectional LSTM (biLSTM) [46] to extract sequential features. Specifically, the convolution feature extracted by the backbone network is first mapped from the three-dimension spatial domain $C \times H \times W$ into a sequence of W vectors,

each having $C \times H$ dimensions. Then, the biLSTM is employed to explore the sequence’s dependencies in two directions, which outputs another sequence with the same length as the input. Finally, the widely-used CTC is employed to decode the sequential feature s_t into label sequence.

4.4 Experiments

4.4.1 Datasets

To evaluate the effectiveness of the proposed methods, we conducted thorough experiments on three widely used license plate recognition datasets, *i.e.*, CCPD-2018 [182], CCPD-Green [182], and CLPD [196].

CCPD-2018 [182] is one of the largest public license plate datasets, which provides almost 300k images that were taken in China. Each image contains a unique LP with detailed annotations, including bounding box coordination, ID, blurring level, *etc.* The entire dataset has been separated into 7 subsets based on different environmental conditions and degrees of tilt. Each split contains 10~20k images except the *base* split which has approximately 200k samples. Following [182], 100k images in the *base* set are used for training while the other half, as well as the rest splits¹ (*DB, FN, Rotate, Weather, and Challenge*) are used for testing.

CCPD-Green [182] is an extended subset of CCPD-2018, which contains a total of 11,776 images, including 5,769 for training, 1,001 for validation, and 5006 for testing. Unlike CCPD-2018, this subset only contains green LPs for new energy vehicles.

AOLP [55] offers 2049 license plate photos that were taken in Taiwan. The images are split into 3 subsets: Access Control (AC) with 680 images, Traffic Law Enforcement (LE) with 757 images, and Road Patrol (RP) with 611 images. We follow the same training settings as presented in previous works [86, 175, 196], in which samples from different sub-datasets are used for training and testing, respectively. Besides, we also generate 10k synthetic samples for each fold by the proposed TLPNet.

CLPD [196] contains 1200 real images, which cover various photographing conditions with different types of LPs (blue, green, yellow, *etc.*). Different from the CCPD dataset, CLPD is a more balanced benchmark in terms of the distribution of region code, as the pictures were collected from all provinces in

¹The CCPD splits were updated after publication, to conduct a fair comparison with existing approaches, we adopt the original splits.



FIGURE 4.6. Comparison of real images and synthetic data generated by different algorithms.

mainland China. Due to the small volume, we only use the CLPD dataset as a test bed to evaluate the generalization ability of recognition models.

4.4.2 Implementation Details

As this chapter focuses on the recognition of license plates, we trained an off-the-shelf YOLO [3] detector on the training split of each dataset to obtain the bounding boxes of LPs. The detector achieves $AP@0.5 = 99.88$ on the CCPD validation set. We follow the identical evaluation protocols as used in [182], *i.e.*, the prediction result is correct only if the IoU between predicted and ground-truth bounding-boxes is greater than 0.6 and all characters of the LP ID are correctly recognized. Each experiment was conducted on a single NVIDIA Tesla V100-SXM2-32GB GPU.

To train the proposed Text to License Plate Network (TLPNet), 6,000 images, including 3,000 starts with “皖” and the other half plates are of other provinces, from the CCPD-2018 [182] training split are randomly sampled, on which the TLPNet is trained for 70 epochs. Online data augmentation strategies, including random cropping, rotation, and blurring, are adopted for enlarging the training samples. Instead of using complex methods such as CLIP [136] to filter low-quality images, we directly set a threshold based on the file size to select qualified images.

4.4.3 Comparison with real data

Experiments on different numbers of real photos and images synthesized by the proposed methods are conducted to assess the quality of generated data. Specifically, the generator was trained within 6k real images. Meanwhile, the recognizers are trained using identical hyper-parameters for 100 epochs with a batch size of 256 and evaluated on the full CCPD2018 testing split. Table 4.2 shows that the model trained on the real samples still achieved better results using the same number of training images. However, the model trained with synthetic data can also obtain comparable performance when the training data is increased. For example, the model trained with 30k synthetic images achieved 94.2 and 88.5 on *Base* and *Weather*, respectively, while the 5k real-image-trained model obtained 95.3 and 90.5. Nonetheless, considering the difficulty of data acquisition between real images and synthetically generated data, it is much easier and cheaper to access a massive amount of synthetic data. Therefore, we synthesized a large scale of images (~ 1 million) and tested the recognizers that were trained with a different number of synthetic samples. Table 4.2 shows that a model trained with synthetic data which $10\times$ number of real images can achieve comparable performance. For example, the model trained with 50k synthetic images achieved an 87.5 overall score, while its counterpart, the model trained with 5k real data, obtained 88.2 overall precision. Similarly, the models trained on 50k real images and 500k synthesized LPs achieved 96.8 and 96.4, respectively. Furthermore, the recognizer trained on 1 million synthetic data obtained competitive results to the model trained on the 100k training data of the CCPD dataset (97.4 vs. 97.7), while the model trained on a combination of 100k TLPNet generated samples and 20k OpenCV generated samples, the performance even outperform the models trained on real data.

In addition, as discussed in Sec. 4.1, the majority of samples in the CCPD dataset share the same province code “皖”, which also induces the lack of training samples for the proposed TLPNet. To further evaluate the quality of the synthesized images with fewer training samples, we built a split that excludes all of the samples that contain “皖” in the standard test splits (7467 out of 180k), termed *minor*. Compared to other splits, both models trained on real data and synthesized images perform worse on *minor* due to the lack of training samples. Nevertheless, as shown in Table 4.2, the baseline recognizer still gains significant improvements on the *minor* split while adding more synthetic images. Especially, the model trained on 500k TLPNet generated images even beat the model trained on 50k real data on this split (79.2 vs. 72.5). However,

TABLE 4.2. Performance comparison between models that are trained on real images and synthetic data on CCPD2018 test set. Minor* is a subset that excludes samples started with 皖.

#Im	All	Base	DB	FN	Rot.	Tilt	Weat.	Chall.	Minor*
Real data									
5k	88.2	95.3	91.9	90.3	40.8	73.3	90.5	66.1	51.0
10k	91.7	97.0	93.1	93.0	62.2	82.1	93.2	71.0	59.7
30k	94.4	98.2	96.1	95.6	74.9	87.6	94.8	77.2	72.5
50k	96.8	98.9	96.9	97.6	90.1	94.0	96.9	83.0	76.9
100k	97.7	99.5	98.6	98.7	90.5	95.1	97.8	85.8	78.8
TLPNet synthesized data									
5k	31.4	40.3	20.3	28.3	4.7	17.4	35.4	7.2	10.1
10k	78.7	88.0	66.6	80.4	53.0	72.0	79.0	37.9	21.9
30k	88.3	94.2	81.4	89.8	75.6	84.7	88.5	55.3	42.5
50k	90.3	95.4	84.9	91.5	83.3	87.5	89.3	58.2	53.2
100k	92.7	97.0	88.8	94.4	84.5	89.7	92.2	65.6	67.9
200k	94.9	98.1	92.1	96.3	89.6	93.0	94.0	73.4	68.5
300k	95.6	98.5	93.4	96.8	91.3	94.4	94.7	75.1	73.6
500k	96.4	98.8	95.4	97.2	92.0	94.8	95.4	80.0	79.2
1m	97.4	99.3	96.5	98.0	93.1	95.7	96.5	82.3	80.2
TLPNet (1m) + OpenCV (200k)									
1.2m	97.8	99.5	97.5	98.6	94.7	95.7	97.6	86.7	82.3

TABLE 4.3. Comparison of recognition performance using ground-truth and detection bounding box.

Type	Overall	Base	DB	FN	Rot.	Tilt	Weat.	Chall.
GT-box	97.9	99.6	98.6	98.8	91.5	95.5	97.9	86.0
Detection	97.7	99.5	98.6	98.7	90.5	95.1	97.8	85.8

the TLPNet was trained only on 6k real images with annotations. In contrast, the recognizer trained on 5k real images can only achieve 51.0 on minority split, while the TLPNet significantly lifts the performance up to 80.2. This suggests that although the training data is insufficient, TLPNet still generates high-quality images that can provide effective supervision signals.

Besides, to eliminate the possible impact introduced by detection results, Table 4.3 compares the performance of the baseline recognizer based on ground truth, and YOLO predicted bounding box. As shown in the table, the ground-truth bounding box only improves 0.2 overall performance compared to the detected box, which suggests that the impact of detectors is limited.

4.4.4 Comparison with other text synthesis approaches

To compare the quality of TLPNet synthesized images with other methods (see Fig. 4.6), we pre-trained the baseline recognizer on the data generated by several other widely used synthesis techniques designed for both generic text and LP recognition. Then we evaluate these models on CCPD-2018 and CCPD-Green, respectively, to show their effectiveness. For a fair comparison, all models are trained within identical settings unless otherwise specified.

TABLE 4.4. Comparison of performance using different data synthesis methods.

Model	Pre-training	Fine-tuning	Total Images		CCPD-2018				Test Set			
			#Syn.	#Real	Overall	Base	DB	FN	Rotate	Tilt	Weat.	Chall.
Real-C18	None	CCPD-2018	0	100k	97.7	99.5	98.6	98.7	90.5	95.1	97.8	85.8
Real-CG+C18	CCPD-Green	CCPD-2018	0	105k	97.8	99.5	98.8	98.6	90.5	95.0	97.9	86.3
			Synthetic Data Only Training									
Plain	None	Plain-Syn	200k	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
OpenCV	None	OpenCV-Syn	200k	0	59.5	68.1	54.9	62.1	7.5	46.6	73.8	28.9
CycleGAN	None	CycleGAN-Syn	200k	0	47.0	53.1	43.6	47.8	4.8	40.3	67.7	18.3
TLPNet	None	TLP-Syn	200k	0	94.9	98.1	92.1	96.3	89.6	93.0	94.0	73.4
Combination	None	TLP-Syn+OpenCV	1.2m	0	97.8	99.5	97.5	98.6	94.7	95.7	97.6	86.7
			Real + Synthetic Data Training									
Real+MJSynth	MJSynth [59]	CCPD-2018	7m	100k	97.8	99.4	98.7	98.4	92.1	95.1	97.5	86.9
Real+SynthText	SynthText [49]	CCPD-2018	4m	100k	97.6	99.4	98.5	98.6	90.8	94.2	97.7	86.2
Real+Plain	Plain-Syn [2]	CCPD-2018	200k	100k	96.5	99.1	98.1	97.8	85.0	91.5	96.6	81.0
Real+OpenCV	OpenCV-Syn [1]	CCPD-2018	200k	100k	98.1	99.7	98.9	98.6	92.1	95.0	98.2	87.7
Real+CycleGAN	CycleGAN+Syn [207]	CCPD-2018	200k	100k	98.1	99.7	99.1	98.9	92.1	96.0	98.4	86.8
Real+TLPNet	TLP-Syn	CCPD-2018	200k	100k	98.5	99.7	99.2	99.2	94.3	96.5	98.8	88.9
Real+Combination	TLP-Syn+OpenCV	CCPD-2018	1.2m	100k	99.2	99.8	99.5	99.5	98.3	98.1	99.2	93.3

MjSynth: The NeurIPS2014 [59], *a.k.a.* MjSynth, is a widely used pre-training dataset for text recognition, which has 9 million synthetic images covering 90k English words. As this dataset contains too many samples, we only pre-trained our baseline on the training split (~ 7 million images) for 5 epochs to save time, which still costs more than 50 training hours on a single Nvidia Tesla V100 card.

SynthText: Similar to [59], the CVPR2016 [49], *a.k.a.* SynthText, is another large synthetic dataset designed for text recognition. The difference between MjSynth [59] and SynthText [49] is that each sample in the MjSynth is only a patch containing one single English word with synthetic background, while SynthText placed multiple word instances into a natural photo based on the background layout. Thus, SynthText provides both bounding boxes and text annotations. As this chapter only focuses on the recognition part, we cropped ~ 4 million patches from the original synthetic scene images in SynthText, and each image patch contains an English word with a natural background. We also pre-trained the baseline for 5 epochs on this dataset.

Plain-Syn: Both [59] and [49] are designed for generic text recognition, which only covers English words and numbers. To investigate the influence of the vocabulary, a widely-used Text Recognition Data Generator (TRDG) [2] is employed to generate LP text rather than dictionary words. Specifically, 200k image patches are generated. Each patch contains only one valid LP number with an empty background. Random blurring and skewing are adopted to make the data more diverse. The baseline method was trained on a total of 200k Plain-Syn datasets for 50 epochs.

OpenCV-Syn: As the format of car LP is relatively fixed, a popular synthesis method is to directly render random numbers onto blank LP images, then employ image processing techniques to add noise, distortion, deformation, etc. An off-the-shelf toolkit [1] developed upon OpenCV was employed to generate 200k LPs. The baseline recognizer was then trained on the combination of 200k synthetic LP images for 50 epochs.

CycleGAN-Syn: Some previous works [168, 196] have proved that CycleGAN [207] can be used to transfer the synthetically generated data into photo-realistic images, which reduces the domain gap between real and synthetic data. In specific, a CycleGAN model was trained to transfer the 200k images in the OpenCV-syn into the photo-realistic domain. The baseline recognizer was then trained on the new images for 50 epochs.

TLP-Syn (ours): Similarly, we trained a TLPNet model on 6k real images based on the implementation details presented in Sec. 4.4.2. Based on this

model, 200k LP images are synthesized. The baseline recognizers were trained within the identical hyper-parameters used in other baselines for 50 epochs. Examples of TLP-Syn can be found in Fig. 4.8.

As shown in Table 4.4, the baseline recognizer achieves an overall performance of 97.7 on the CCPD-2018 test split without external training data. While adding the real CCPD-Green data for pre-training, the performance slightly improved to 97.8. To assess the quality of samples that were synthesized by different approaches, we compare the proposed TLPNet with five other synthesis methods, MJSynth [59], SynthText [49], Plain [2], OpenCV [1], and CycleGAN [207]. We evaluate these methods from two aspects, *i.e.*, synthetic data only training and synthetic data pre-training with real data fine-tuning. It should be noted that the MJSynth and SynthText do not cover any LP samples; thus, they are not applicable for synthetic-data-only training evaluation.

Synth-Only: Table 4.4 shows that the baseline trained on Plain-Syn failed to distinguish authentic images, which suggests that the background, font, and style are essential for training the LP recognizers. The OpenCV and CycleGAN model achieved 59.5 and 47.0 overall accuracies on the CCPD-2018 test split, respectively. Surprisingly, the images adapted from synthesized images to the photo-realistic domain by the CycleGAN did not help improve the model performance but even induced a significant precision drop. With identical training settings, the baseline recognizer trained on the images generated by the proposed TLPNet achieved a 94.9 overall score on the CCPD-2018, which shows comparable performance to the models trained on real images. Moreover, to further explore the capability of synthesized images, a combination of samples generated by two different approaches is built, which includes 1 million images from TLP-Syn and 200k images from OpenCV-Syn. This setting further boosts the baseline recognizer to 97.8, which has already beaten the real data-only trained model Real-C18 (97.8 vs. 97.7), demonstrating that training on a combination of synthetically generated data can obtain competitive results to the real images.

Synth+Real: We also conduct experiments to evaluate the fine-tuned models. Formally, each model pre-trained on the synthetic data is further fine-tuned on the CCPD-2018 training set for 50 epochs. As shown in Table 4.4, both models were pre-trained on the generic OCR synthetic data MjSynth and SynthText, which obtained similar accuracy of 97.8 and 97.6, respectively. Although the recognizer was pre-trained on millions of images, it does not gain improvements compared to the real-data-only trained models. A possible reason is that both the image and the dictionary of these two synthetic datasets are quite different

TABLE 4.6. Impact of the number of labeled data while training the generator.

VAE Stage			Transformer Stage		
#Unlabeled	Acc.	NED	#Labeled	Acc.	NED
10,000	81.9	95.8	1,000	82.5	95.6
30,000	84.1	96.5	3,000	83.9	96.2
50,000	84.1	96.5	5,000	84.0	96.2
100,000	84.2	96.5	10,000	84.3	96.3

TABLE 4.7. Performance of the proposed methods on the CCPD dataset.

Model	Pre-train	Fine-tune	Accuracy
Real	None	CCPD-Green	82.5
Real+Real	CCPD-2018	CCPD-Green	90.5
Synth+Real	TLP-Syn	CCPD-Green	91.0

from the LP recognition task, where a large domain gap exists and prevents the improvements. Plain can be regarded as a particular case of MjSynth, which replaces the preset English dictionary with LP text. However, the combination of real data and Plain synthesis encounters a significant accuracy drop, even worse than the train-from-scratch model (96.5 *vs.* 97.7). This suggests that pre-training the recognizer on a small data set with a large domain gap can bring a bad parameter initialization to the network, harming the final performance. Although OpenCV-Syn performs better than CycleGAN-Syn on the synthetic-data-only training phase, the recognizers fine-tuned upon these pre-trained models obtain a similar overall performance of 98.1 on the CCPD-2018. When it fine-tunes the proposed TLP-Syn, the recognizer further obtains 0.4 improvements in the overall score. The combination setting that involves 1 million TLP-Syn, and 200k OpenCV-Syn images significantly improved the recognition performance compared to the real-data-only models, which suggests that the massive synthetic data helps the model learn discriminative features. Benefiting from the varieties of external synthesized images, the model has achieved better performance on challenging samples such as rotated and tilted LPs without introducing complex mechanisms like 2D attention.

TABLE 4.8. Comparison of performance between ours and the state-of-the-art methods on the CLPD [196] dataset.

Model	Acc	Acc w/oRC
Masood et al. (2017) [200]	-	85.2
Xu et al. (2018) [182]	66.5	78.9
Zhang et al. (CCPD2018 Only) (2020) [196]	70.8	86.1
Zhang et al. (CycleGAN+CCPD2018) (2020) [196]	76.8	87.6
Wang et al. [174]	89.8	95.3
Ours (TLP-Synthetic Only)	85.3	91.0
Ours (CCPD Only)	78.0	90.7
Ours (TLP-Synthetic+CCPD)	91.2	95.6

TABLE 4.9. Comparison of performance between ours and the state-of-the-art methods on the AOLP [196] dataset.

Model	Overall	AC	LE	RP
#Images	(2049)	(681)	(757)	(611)
Li et al. [86]	94.5	95.3	96.6	83.7
Zhang et al. [196]	96.1	97.3	98.3	91.9
Zou et al. [196]	97.8	99.3	98.7	95.1
Wang et al. [174]	99.7	99.4	99.9	99.7
Zou et al. [209]	96.5	96.3	97.9	94.9
Ours (TLP-Synthetic+AOLP)	99.6	99.5	99.6	99.7



FIGURE 4.7. Synthetic LPs that are generated by our methods. Compared to the existing methods that directly render digits and characters on a blank template, the proposed methods learn from existing data to generate more realistic photos. First row: generated LP with less tilt and rotation. Second row: generated LP with large tilt.

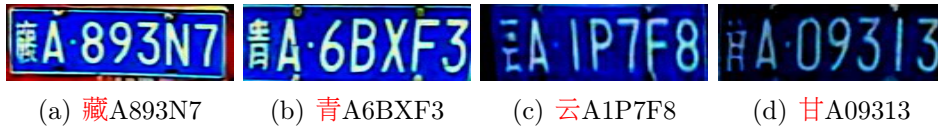


FIGURE 4.8. Failure cases of generated LPs. As the available training data of these provinces is very limited (*e.g.* 藏 (7); 青 (8); 云 (10); 甘 (10)), the proposed method failed to generate high resolution province code.



FIGURE 4.9. Examples of the incorrect recognition results. The samples from the three rows are from CCPD-2018, CCPD-Green, and CLPD, respectively.

4.4.5 Ablation Study

As introduced previously, unlabeled images are used at the VAE training stage, while labeled instances are used at the transformer training stage. Ablation studies are conducted to reveal the impact of the number of labeled and unlabeled samples on training the LP generator. Each variant is employed to generate 10k LPs to evaluate the image generation quality. Then, the generated samples are further augmented to 50k images for training the OCR models. Finally, the OCR models that were trained on synthetic data only are evaluated on the CCPD-2018 under both accuracy and Normalized Edit Distance (NED).

Unlabeled Data: Unlabeled data can be easily accessed, which makes training the models on massive unlabeled data possible. Ablations under 4 different settings are conducted to assess if the unlabeled data can improve the quality of generated images. Specifically, the VAEs have trained with 10k, 30k, 50k, and 100k unlabeled images. Then, each VAE is further used to train the LP generator based on identical 10k labeled data. As shown in the left of Table 4.6, increasing the number of training samples can not bring further performance improvements. This suggests that the VAE part can easily get to convergence

within limited data.

Labeled Data: It is well-studied that increasing the number of training samples can effectively improve the model performance. However, it is more expensive to acquire labeled images compared to unlabeled data. Therefore, ablation studies are conducted to explore enhancing the generation quality with less labeled data to balance the trade-off between cost and precision. In specific, each variant transformer was trained on an identical VAE with 1k, 3k, 5k, and 10k labeled images. As shown in the right of Table 4.6, the quality of generated images is not very sensitive to the number of training samples. Even training on only 1k images can enable the transformer to generate high-quality samples.

4.4.6 Results on CCPD-2018

To evaluate the effectiveness of the proposed pipelines, we compared the performance with other state-of-the-art methods [86, 95, 108, 139, 141, 170, 182, 195, 196, 202] on CCPD-2018 benchmark.

As shown in Table 4.5, our baseline recognizer has already surpassed most existing methods when only training on the synthetic data generated by the TLPNet. Meanwhile, our model achieved the highest performance on *FN*, and the second-highest performance on *Weather*, *Rotate*, and *Base* splits, only slightly worse than [174] which employed a much more sophisticated network. Even for the *Challenge* split that contains the most complex cases, our baseline model still obtains 86.7, outperforming all state-of-the-art approaches except [174, 196].

Like all other methods used real or a combination of real and synthesized images for training, we also fine-tuned the Real+Combination model on the CCPD training split for 50 epochs. As shown in Table 4.5, our algorithm outperforms all other methods in terms of the overall score and most subsets. The only exception is that both [196] and [108] perform slightly better than us on *Rotate* and *Tilt*. The reason might be that both methods [108, 196] employed stronger image rectification techniques, such as 2D attention and rectified attention network, which enables them to obtain better performance on rotated images. However, our algorithm can also benefit from such modules, and the precision is expected to be further boosted on these two subsets. Besides, the increment on *Challenge* split is even obvious; our method achieved an accuracy of 93.3. This is because the TLPNet synthesized many very challenging samples, enabling the baseline recognizer to learn discriminative features. We show some incorrect recognition results in Fig. 4.9.

4.4.7 Results on Extensive Benchmarks

To evaluate the generalization ability of the proposed methods. We further conduct experiments on CCPD-Green [182], CLPD, and AOLP [196].

4.4.7.1 CCPD-Green

For CCPD-Green, we compare our models trained under three different settings: 1) Real model trained on the CCPD-Green training split from scratch; 2) Real+Real model fine-tuned from CCPD-2018 pre-trained model, and 3) Synth+Real model fine-tuned from 1.2m TLPNet synthesized images. As shown in Table 4.7, pre-training the model on CCPD-2018 significantly improved the accuracy from 82.5 to 90.5. Furthermore, the Synth+Real model pre-trained on TLP-Syn even outperforms the Real+Real model, achieving 91.0 accuracy, which suggests that the TLPNet synthesized images can replace the real images for pre-training the recognizers.

4.4.7.2 CLPD

Following [196], we only use CLPD as the test dataset. The models trained for CCPD are directly evaluated as the baseline recognizers on CLPD without further fine-tuning. The TLP-Synthetic only model was trained on the combination of 1.2 million synthetic data; the CCPD2018-only model was trained on the training split of CCPD2018; the last one was pre-trained on TLP-Synthetic data and then fine-tuned on the CCPD2018 training set. The performance was calculated from two aspects, with and without region code (Chinese character) considered. As shown in Table 4.8, our model has already outperformed the real-data trained model of [196] when only trained on the images synthesized by the proposed TLPNet, which shows the generalization ability of the proposed methods. Furthermore, the full model achieves an accuracy of 95.6 without region code, outperforming other state-of-the-art models.

4.4.7.3 AOLP

To further explore the generalization ability of the proposed methods. Table 4.9 compares the proposed methods with the state-of-the-art methods on the AOLP dataset, where we achieve a competitive performance of 99.6 overall accuracies.

4.5 Conclusion

This chapter presents a Text to License Plate Network (TLPNet), which converts text strings to high-quality license plate training samples. TLPNet is built upon a text-to-image framework, which is composed of three modules, *i.e.*, the discrete variational auto-encoder that compresses photos into image tokens; the embedding layer that converts LP numbers into text tokens; and the decoder that can autoregressively model the joint feature of both image and text tokens. Compared to real images, it is much easier and cheaper to access massive synthetic samples using our approach. Therefore, to verify the effectiveness of the proposed TLPNet, we introduce SynthLP, which is a synthetic LP dataset containing 1 million samples synthesized by the TLPNet. Different from previous synthesis methods, which generate low-quality images that can only be used for pre-training the recognizers, the model trained on the SynthLP achieved competitive performance to its counterpart trained on real images from the CCPD-2018 training split. Extensive experiment results show that the TLPNet can generate high-quality LP images that can partially replace real training samples. State-of-the-art performance and exhausted comparisons with previous methods demonstrate the great potential values of our method.

Chapter 5

On the General Value of Evidence, and Bilingual Scene-text Visual Question Answering

Statement of Authorship

Title of Paper	On the general value of evidence, and bilingual scene text visual question and answering
Publication Status	<input checked="" type="checkbox"/> Published <input type="checkbox"/> Accepted for Publication <input type="checkbox"/> Submitted for Publication <input type="checkbox"/> Unpublished and Unsubmitted work written in manuscript style
Publication Details	IEEE Conference on Computer Vision and Pattern Recognition, 10126-10135

Principal Author

Name of Principal Author (Candidate)	Xinyu Wang		
Contribution to the Paper	Proposed the ideas, conducted experiments and draft the manuscript of the paper.		
Overall percentage (%)	70%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	19/Jan/2023

Co-Author Contributions

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

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

Name of Co-Author	Yuliang Liu		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	Jan/18/2023

Name of Co-Author	Chunhua Shen		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	Jan/18/2023

Name of Co-Author	Chun Chet Ng		
-------------------	--------------	--	--

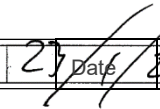
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	Jan/19/2023

Name of Co-Author	Canjie Luo		
Contribution to the Paper	Conduct an experiment.		
Signature		Date	Jan/19/2023

Name of Co-Author	Lianwen Jin		
Contribution to the Paper			
Signature		Date	Jan/19/2023

Name of Co-Author	Chee Seng Chan		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	Jan/19/2023

Name of Co-Author	Anton van den Hengel		
Contribution to the Paper	Discussion and writing revision.		
Signature		Date	Jan/23/2023

Signature:  Date: 23/1/23

5.1 Introduction

The fact that Visual Questions Answering [8] methods are able to answer natural language questions that relate to a wide variety of image contents has been an incredible development. The limitations of existing methods, and particularly their tendency to focus on spurious correlations in the data, have been repeatedly identified (see [4, 43, 65], for example). This is visible in the tendency of methods to answer questions on the basis of text alone. The answer to ‘How many’ questions, for instance, is predominantly ‘Two’.

Focusing on coincidental correlations in the data represents a failure to generalize. These correlations are not stable across datasets, meaning that once the test data moves beyond the distribution of the training set, the correlations fail to hold, and methods that exploit them fail to work. The underlying reasoning, in contrast, is stable across datasets. Encouraging VQA methods to reason about the image content is thus critical to achieving methods that generalize.

One of the underlying problems with encouraging VQA methods to generalize has been that it is impossible to tell whether a method arrived at the right answer for the right reasons. An answer is equally correct whether it results from analysis of the underlying reasoning or through exploiting a coincidental correlation in the data. A series of works have developed more sophisticated measures of performance for vision and language problems [7, 43, 186], and this work falls in this category. What distinguishes this approach is that it uses image-based grounding to encourage generalization, despite the fact that it is not actually required to achieve the desired task.

We propose here an approach to measure VQA performance that encourages generalization by demanding that the algorithm justifies its reasoning (see Figure 5.1). Previous methods have applied the same rationale but suffered because the form in which the reason must be provided is constrictive [166, 177]. We show here that it is possible instead to evaluate reasoning by only requiring a method to provide a relatively simple indication of which area of the image it has based its answer on. If the method provides the correct answer and the correct image region, then it is likely that it has employed the right reasoning. Using image regions, or more accurately bounding boxes, as an evaluation metric also has the advantage that Intersection-over-Union (IoU) measures are well understood in the field.

The version of the VQA problem that we apply this approach to is Scene Text VQA. Several recent works [13, 151] have revealed that current VQA models perform poorly on text VQA datasets, so it represents a compelling challenge



FIGURE 5.1. Requiring that vision-and-language methods provide evidence for their decisions encourages the development of approaches that depend on reasoning and thus that are better able to generalize to new situations. It also helps to build up confidence in the provided answer.

falling within the existing framework. The various forms of text VQA problem are also of great practical importance, because text represents a critical cue to understand the content of an image. More than this, text VQA problems are typically less susceptible to solve through exploiting coincidental correlations in the data.

A variety of text-based VQA datasets [13, 72, 119, 151] have been proposed. However, there is still a significant gap between current algorithm performance and that required to support practical applications [13, 119, 151]. Another motivating factor in selecting text-based VQA rather than the generic version of the problem is that the text-based version of the problem is less susceptible to n-way classification over a fixed vocabulary. This is due to the fact that the range of text appearing in images is quite broad. The classification-based approach has repeatedly been shown to be susceptible to overfitting [4, 43]. Text-based VQA requires the development of alternative approaches, some of which will hopefully generalize.

Figure 5.2 depicts some of the challenges with existing scene-text based VQA system. For example, Figure 5.2(a) is a sample question that can be answered without reference to any textual content; while the question in Figure 5.2(b) could have more than one correct answer; the question in Figure 5.2(c) requires prior knowledge to answer; and finally in Figure 5.2(d), the answer can not be

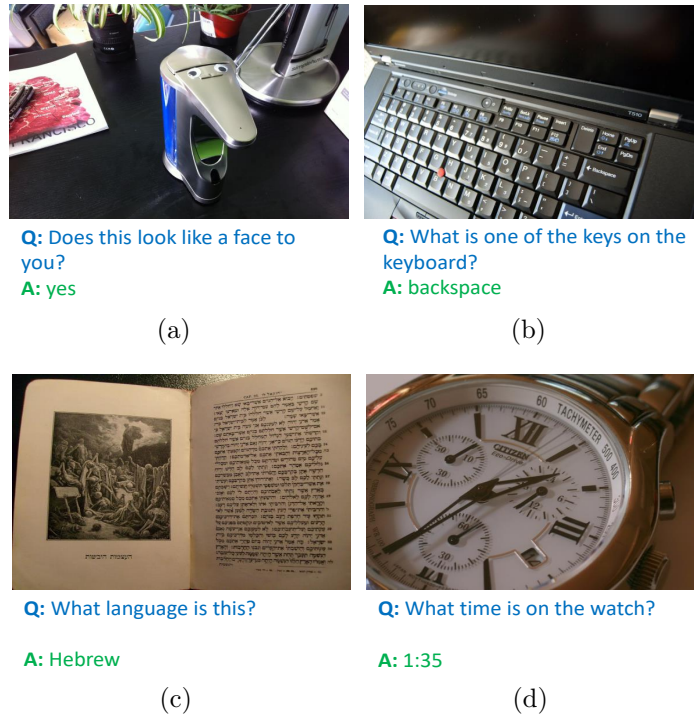


FIGURE 5.2. Some example images and QA pairs from the Text-VQA proposed in [151]. Four different types of issues are shown: (a) questions that can be answered without reading image text; (b) questions that have more than one correct answer; (c) questions that require a large amount of external knowledge to answer; (d) questions that require skills that cannot be learned from the training data alone.

Q: What state does this team play for?



FIGURE 5.3. A comparison of conventional (LoRRA [151]), and evidence-based VQA methods.

obtained directly from the text in the image, but require other skills.

Empirical results presented in Figure 5.3 demonstrate that current VQA approaches rely heavily on a pre-defined answer space constructed by analysis of the answers in the training set, and thus limiting generalization. As shown

in Figure 5.3(b), their dependence on superficial image features can render conventional VQA methods sensitive to image modifications that do not change the semantics. Figure 5.3(c) and 5.3(d) demonstrate their propensity to generate an answer even when the required information is not present.

Text-VQA [151] employed the generic VQA accuracy as the performance metric, while ST-VQA [13] used a soft score metric inspired by the optical character recognition community. Both of these metrics are results-oriented, which means that a prediction is deemed correct if it is identical to the ground-truth. They do not assess the reasoning process. Such classification-based VQA models are able to achieve impressive performance but they are prone to overfit a fixed answer space and generalize poorly to other datasets.

To address these issues, we propose a new scene-text based VQA dataset called ‘Evidence-based Scene Text Visual Question Answering’ (EST-VQA). Based on this, three tasks namely *cross language challenge*, *localization challenge* and *traditional challenge* are introduced to motivate the creation of solutions with practical value from various aspects. Also, a series of baseline experiments were conducted to establish a lower bound for these three challenges. The main contributions of this paper are outlined as follows:

- **Dataset:** The EST-VQA dataset provides questions, images and answers, but also a bounding box for each question that indicates the area of the image that informs the answer. We refer to such bounding boxes as *evidence*. The dataset is intended to enable the development of text VQA methods that are closer to the levels of performance required by practical applications, but also to encourage the development of general VQA methods that generalize.
- **Evaluation Metric:** We introduce an Evidence-based Evaluation (EvE) metric, which will require a VQA model to provide evidence to support the predicted answer. For this purpose, a new VQA model is also proposed. Under this new metric, it is anticipated that it will be much more difficult for naive classification models to achieve inflated performance.
- **Bilingual:** To the best of our knowledge, the proposed EST-VQA is the first bilingual scene text VQA (ST-VQA) dataset that includes both English and Chinese question and answer pairs. The fact that the proposed dataset embodies questions in two languages further rewards methods that generalize well. It is more difficult for a method to exploit superficial correlations in questions expressed in multiple languages. The languages chosen are also particularly grammatically distinct, and reflect culturally

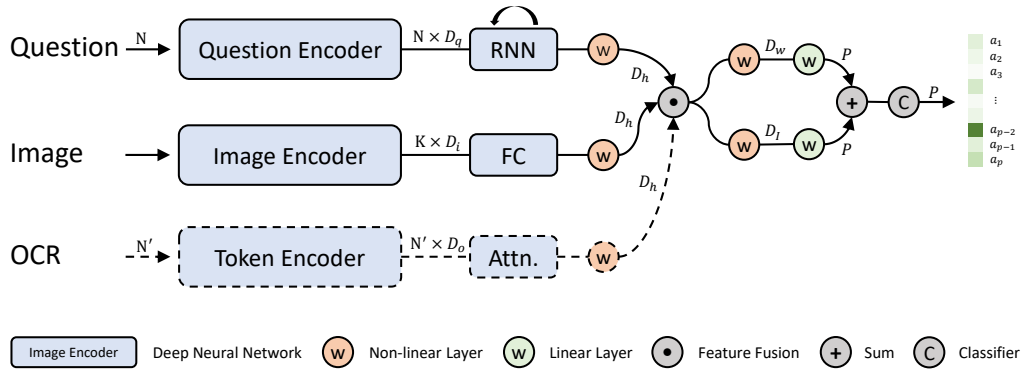


FIGURE 5.4. Illustration of the mainstream VQA models. D_q , D_i , D_o and D_h are the dimensions of the word embedding, image feature, OCR token embedding and hidden vector representations respectively. N , N' and P indicate question length, number of OCR tokens and answer space. Blocks with dashed lines are optional modules used for text-based VQA.

distinct populations, which leads to different question statistics, and further encourages generalization.

5.1.1 Related Work

Visual Question Answering has gained significant attention recently, partly because it seems so unlikely that a method might be capable of answering all possible questions about all possible images [8, 113]. Readers are encouraged to refer to [67, 178] for a complete overview. Due to space constraint, this section only reviews the most relevant works to this paper, *i.e.*, text-based VQA.

5.1.2 Text-based VQA

In contrast to generic VQA datasets [67, 178], text-based VQA datasets pay more attention to text-related questions where a VQA model is required to read and understand textual content in an image. In [151], the authors proposed a dataset and baseline model, called Text-VQA and LoRRA respectively. LoRRA follows the structure of mainstream VQA models (see Figure 5.4) where image features and word embedding are fused to train a classifier. Later, two other similar datasets were introduced, *i.e.*, ST-VQA [13] and OCR-VQA [119]. All these three datasets provide images with text-related question-and-answer pairs. However, there are several important differences between them, as well as our proposed dataset:

Diversity: Table 5.1 shows the size and image sources of existing datasets and our dataset. Both of the Text-VQA [151] and OCR-VQA [119] images

model is ans , then the score for a single sample is calculated as:

$$s_v(ans) = \min\left\{\frac{\#\text{humans that said } ans}{3}, 1\right\} \quad (5.1)$$

where $\#$ indicates the number of human-annotated labels that are identical to the predicted answer. This metric is robust against the incorrect answers given by some annotators. However, it is clear that only 4 discrete scores would appear, *i.e.*, $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$. In [13], Levenshtein distance [82] was proposed to softly penalize a mistake. Given the predicted answer ans and ground-truth label gt , then the normalized Levenshtein similarity score s_l is given as:

$$s_l(ans, gt) = s_l(ans, gt) = \begin{cases} 1 - NL(ans, gt), & NL(ans, gt) < \tau \\ 0, & NL(ans, gt) \geq \tau \end{cases} \quad (5.2)$$

where τ is a penalty threshold, and NL is the normalized Levenshtein distance between ground-truth and prediction.

5.2 Proposed Dataset: EST-VQA

A fundamental hypothesis in EST-VQA dataset is that a VQA model should answer a question correctly based on the textual content in an image. Therefore, we separate our scene text VQA tasks into two parts, *i.e.*, 1) text spotting and 2) question answering. In this section, we describe the process of building the EST-VQA dataset. Also, we will detail the evidence-based evaluation metric and the new tasks for EST-VQA dataset.

5.2.1 Data Collection

Images: As the EST-VQA dataset is designed for scene text VQA tasks, we collected a total of 20,757 images from publicly available scene text detection and recognition datasets. Specifically, images annotated with English questions and answers are obtained from Total-Text [26], ICDAR 2013 [70], ICDAR 2015 [71], CTW1500 [100], MLT [124], and COCO Text [161]. Whereas, images with Chinese questions and answers are collected from LSVT [156]. All the images originated from these scene text datasets are comprised of daily scenes that include both indoor and outdoor settings.

Questions and Answers: The proposed EST-VQA dataset consists of 15,056 English questions and 13,006 Chinese questions. The question and answer pairs could be formed in cross-language *e.g.*, an English question queries the name

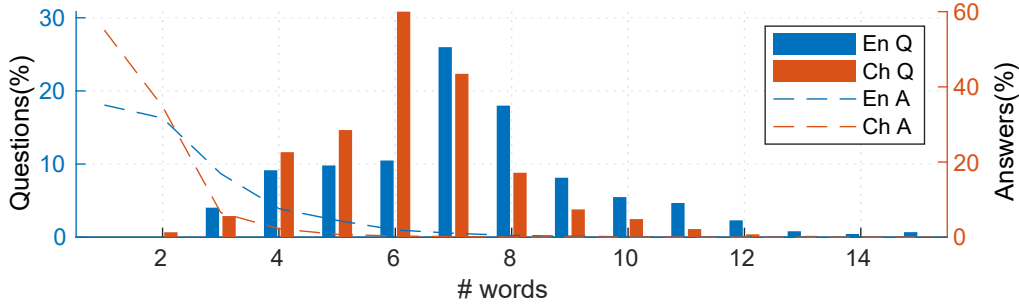


FIGURE 5.6. Percentage of question and answer length in EST-VQA dataset. Questions are tokenized by words. En and Ch stand for English and Chinese respectively.

Set	English		Chinese		All	
	# I	# Q	# I	# Q	# I	# Q
Train	11,383	12,638	9,374	10,506	20,757	23,144
Test	2,267	2,514	2,215	2,500	4,482	5,014
Total	13,650	15,152	11,589	13,006	25,239	28,158

TABLE 5.2. Volume of the EST-VQA dataset.

of a Chinese restaurant so that the answer could be a Chinese text and vice versa for Chinese question. For the collection of question-and-answer pairs, annotators were requested to come up with questions that can be answered only by reading texts in the images. In order to avoid the question that does not require reading any text in the image, annotators are enforced to label a corresponding quadrilateral bounding box of the textual answer. The annotated bounding box will then serve as an evidence to support the answer. Moreover, yes/no questions and ambiguous questions that could have multiple correct answers are prohibited. Figure 5.5 shows the common types of questions, it is clear that most of the English questions start with “what”, and follow by ‘is’ and ‘the’. However, the composition of Chinese questions is far more complex than the English questions due to differences in grammar, vocabulary and other characteristics of the Chinese language. Figure 5.6 shows the distribution of the length of questions and answers. Different from English words which can be segmented by space directly, Chinese words are composed of multiple Chinese characters in a continuous sentence. Therefore, we use [155] to tokenize Chinese questions for counting the percentage of question length. From Figure 5.6, it is clear that most of the English and Chinese questions have between 6 to 8 words, and the majority of their answers are of a single word.

In summary, as shown in Table 5.2, 25,239 images and 28,158 QA pairs are separated into 20,757 images with 23,144 questions for the training set and 4,482 images with 5,014 questions for the testing set.

5.2.2 Evidence-based Evaluation (EvE) Metric

We observed an intriguing trend among the classification based approaches for scene text VQA task. That is to say, if the ground-truth answer was included in the pre-generated answer dictionary, a generic VQA model may predict a correct answer without reading the textual content. However, such methods rely heavily on the pre-defined answer pool and so, they are unable to handle questions with out-of-vocabulary answers. Therefore, it is unclear whether such models truly have the capability to understand and reason about the questions or they are merely over-fitting to the fixed answer space. Inspired by this observation, we introduce a new evaluation protocol, named Evidence-based Evaluation (EvE) metric, which will require a VQA model to provide evidence to support the predicted answers. Under this metric, it will be much more difficult for naive classification models to achieve inflated performance.

Generally, EvE metric consists of two steps: a) check the answer; b) check the evidence. In the former, we use the normalized Levenshtein similarity score (see Eq. (5.2)). In the latter, we adopt the widely used IoU metric to determine whether the evidence is sufficient or insufficient. Suppose B_{gt} and B_{det} are the ground-truth and predicted bounding box respectively, then the evidence sufficiency score, E is defined as:

$$E_{\tau}^i = f\left(\frac{B_{gt} \cap B_{det}}{B_{gt} \cup B_{det}}\right) = \begin{cases} \text{Incorrect}, & E = 0 \\ \text{Insufficient}, & 0 < E < \theta \\ \text{Sufficient}, & E \geq \theta \end{cases} \quad (5.3)$$

where $\theta = 0.5$ is a predefined threshold. Under the EvE metric, only *correct* answers with *sufficient* evidence contribute to the final performance s_e (see Figure 5.7) where it is given by:

$$s_e(ans, gt, E) = \begin{cases} s_l, & \text{if } E \text{ sufficient} \\ 0, & \text{else} \end{cases} \quad (5.4)$$

where s_l is the normalized Levenshtein similarity score as defined in Eq. (5.2).

5.2.3 Tasks

Both Text-VQA [151] and OCR-VQA [119] follow the same rules as presented in generic question answering task. Although ST-VQA [13] proposed three tasks, the only difference between each of the tasks is the size of external information

Question: How many milligrams are the Valium 2?



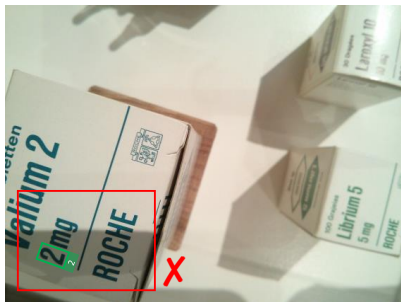
A: '2'

(a) Without Evidence



A: $[[x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4], '2']$

(b) Incorrect Evidence



A: $[[x'_1, y'_1, x'_2, y'_2, x'_3, y'_3, x'_4, y'_4], '2']$

(c) Insufficient Evidence



A: $[[\bar{x}_1, \bar{y}_1, \bar{x}_2, \bar{y}_2, \bar{x}_3, \bar{y}_3, \bar{x}_4, \bar{y}_4], '2']$

(d) Sufficient Evidence

FIGURE 5.7. In EvE metric, evidence in the form of bounding box should be provided as well as the predicted answer. Green and red bounding boxes are ground-truth and predicted evidence respectively. **Incorrect:** (a) answer without evidence; (b) answer with inappropriate evidence; (c) answer with insufficient evidence. **Correct:** (d) answer with appropriate evidence. It is worth mentioning that all of the above answers would be marked as correct in the conventional VQA evaluation metric because all of them give the right answer '2'.

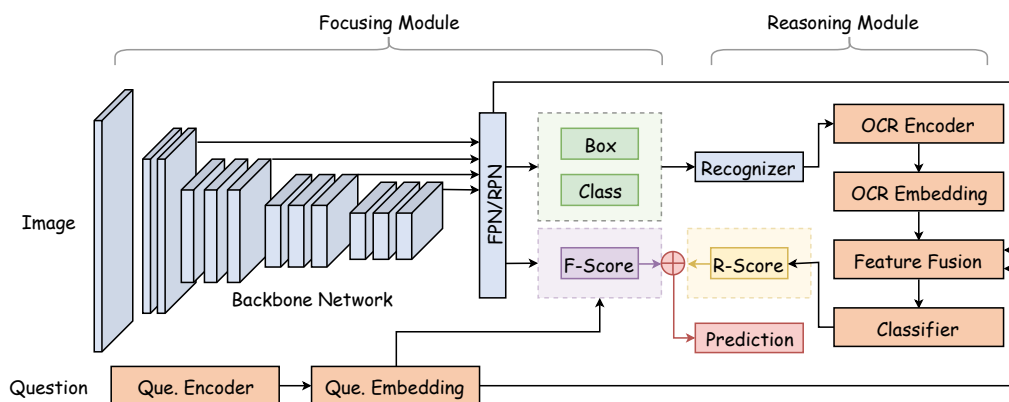


FIGURE 5.8. Overview of the QA R-CNN architecture.

(vocabulary), which is insignificant and unreasonable to properly evaluate the models' full capability. For instance, in the strongly contextualized task, all ground-truth answers are provided in a dictionary for every image with a set of distractors, which makes the VQA model prone to overfit the provided vocabulary. Besides, it becomes more difficult for these models that are trained on a fixed dictionary to generalize to other datasets.

As a result of this, we propose three related tasks namely as *Cross Language Challenge*, *Localization Challenge*, and *Traditional Challenge* that will be detailed next to improve the task diversity. An online evaluation server will be set up for results submission.

- **Cross Language Challenge (CLC):** As the proposed EST-VQA dataset is a bilingual VQA dataset that contains both English and Chinese QA pairs. This challenge aims to explore a model's ability in extracting common knowledge between different languages. Under this challenge, the candidates are requested to submit results predicted by both the monolingual (*English-only*, *Chinese-only*) and bilingual models with an identical framework (*e.g.* network structure) for evaluations. The proposed EvE metric is used to evaluate the model's performance in this challenge.
- **Localization Challenge (LC):** To gain insights into a VQA model, we encourage candidates to train an evidence based VQA model to simultaneously predict the answer and its corresponding bounding box as evidence, instead of simply employing an off-the-shelf OCR system to obtain the OCR tokens. Hence, the main objective of this challenge is to explore the VQA model's ability in understanding the question and locating the correct image space that contains the answers. That is to say, this challenge requires the VQA model to provide the spatial location where an answer will be most likely to appear in an image based on a question. Compared to the full challenge, LC ignores the text recognition error and the difficulties of combining multiple OCR tokens for long answers. IoU between the predicted and ground-truth bounding box is employed as the performance metric for this challenge.
- **Traditional Challenge (TC):** We maintain the traditional VQA challenge that is consistent with the existing VQA datasets in which this challenge does not consider the evidence for the predicted answers. The normalized Levenshtein similarity score between the prediction and ground-truth is employed as the metric for this challenge.

Model	CLC (%)						LC (%)						TC (%)						Δ_r		
	Mono.			Bi.			Mono.			Bi.			Mono.			Bi.					
	En	Ch	Acc	En	Ch	Acc	En	Ch	Acc	En	Ch	Acc	En	Ch	Acc	En	Ch	Acc			
SV UB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	31.1	7.8	31.3	8.9	20.1	-
LV UB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	48.0	16.1	48.3	17.0	32.7	-
OCR UB	33.9	24.5	33.9	24.5	44.1	14.3	29.2	50.0	37.8	43.9	-	-	-	-	-	38.5	28.2	38.5	28.2	33.3	-
Random	4.4	1.1	4.7	1.2	5.1	0.8	3.0	15.1	5.1	10.1	-	-	-	-	-	5.8	1.5	5.9	1.5	3.7	0.81
P[150]+SV	4.3	0.1	4.5	0.1	4.3	0.2	2.3	17.2	1.8	9.5	-	-	-	-	-	8.0	0.7	7.7	0.7	4.2	0.54
P[150]+LV	4.7	0.2	4.4	0.2	4.2	0.3	2.3	17.4	2.4	9.9	-	-	-	-	-	9.2	0.8	8.2	0.6	4.4	0.52
L[151]+SV	8.2	1.2	8.4	2.0	9.6	0.8	5.2	18.0	5.4	11.7	-	-	-	-	-	12.0	2.6	13.2	3.3	8.2	0.63
L[151]+LV	7.7	0.5	6.8	0.7	6.8	0.7	3.8	18.5	3.9	11.2	-	-	-	-	-	12.0	1.6	11.2	1.7	6.5	0.58
QA R-CNN	7.7	1.4	8.8	3.2	10.8	1.1	6.0	18.3	7.3	12.8	-	-	-	-	-	9.6	2.2	10.6	4.0	7.3	0.82
QA R-CNN w/ tricks	7.4	1.5	8.4	2.9	10.3	1.0	5.7	18.3	7.2	12.8	-	-	-	-	-	11.8	7.9	12.7	9.4	11.0	0.52

TABLE 5.3. Quantitative results of the three tasks in EST-VQA dataset. Mono. and Bi. represent monolingual and bilingual model respectively while S and L are short (one word) and long (more than one word) answers. Scores in bold are the best performance across models.

5.3 Baselines and Results

5.3.1 Baseline Methods

This section presents the naive baseline models and two state-of-the-art VQA methods [150, 151] that were employed in the experiments. This helps to show the difficulty of the proposed EST-VQA dataset and the new tasks. The entire EST-VQA dataset is separated into *training* and *testing* sets (see Table 5.2), and 10% data from the *training* set is used for validation.

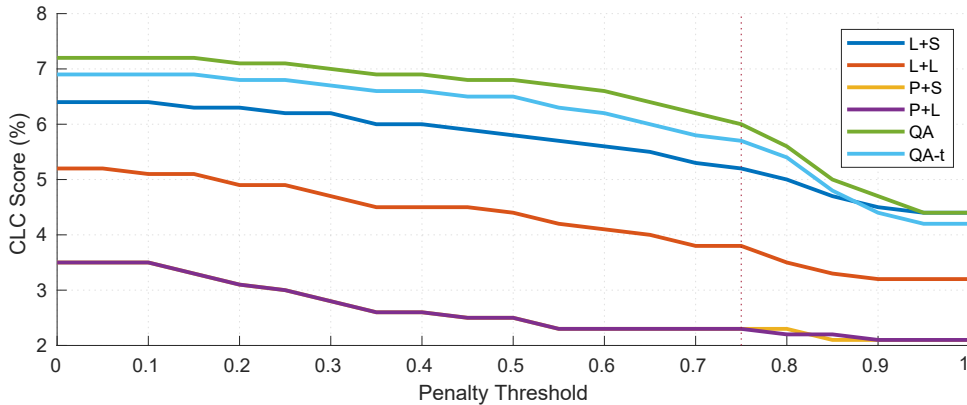
Vocabulary Upper Bound: As both [150] and [151] are classification based, two dictionaries are built under the widely used rules. Specifically, a small vocabulary (SV) is built with 927 English and 365 Chinese answers that appeared more than once in the training set and a Larger Vocabulary (LV) is built with 8,102 English and 8,212 Chinese unique answers. We explore the upper bound accuracy of the pre-generated SV and LV. We assume that answers included in the dictionaries can always be predicted correctly with perfect evidence to calculate the upper bound accuracy.

OCR Upper Bound: Since the traditional VQA models cannot obtain OCR tokens and info directly, we employ the state-of-the-art pre-trained text detection and recognition models [101, 143] to extract OCR bounding boxes and characters. To evaluate the effectiveness of the OCR system, we calculate the OCR upper bound accuracy on the test set. All of the answers and evidence are directly obtained from the OCR results (and suppose the correct one can always be selected), it also considers combinations of up to 4 OCR tokens for multi-word answers.

Random OCR Tokens: To assess arbitrary chance, this baseline returns a random OCR token and its bounding box from the OCR results for each question to obtain random accuracy.

State-of-the-art Approaches: Both state-of-the-art generic [150] and scene text [151] VQA models are employed as baselines to verify the difficulties of the EST-VQA dataset. It is important to note that these methods cannot provide evidence to support their predicted answers. Therefore, we queried the predicted answers from OCR results, *i.e.*, if there are any identical OCR tokens to the predicted answer, then one of the predicted bounding boxes would be randomly selected as evidence, otherwise bounding box of the token which has the smallest normalized Levenshtein distance is selected.

QA R-CNN: It is noteworthy that all of the aforementioned baseline methods cannot simultaneously output the answer and its corresponding bounding box as

FIGURE 5.9. CLC score under different τ

evidence. Therefore, we propose QA R-CNN. Generally, QA R-CNN consists of two parts: Focusing Module (FM) and Reasoning Module (RM) (see Figure 5.8). The core component in FM is a customized Faster R-CNN network trained for text detection task. Compared to the regular Faster R-CNN which only predicts the bounding box and object category, QA R-CNN also outputs a focusing score for each of the bounding boxes. Technically, word embedding of question is first extracted by GloVe [130] for English questions and Word2Vec [89] for Chinese questions. Then, the embedding is fed into LSTM layers to obtain question features. Following this, both question and image features are concatenated to classify the bounding box into answer area and non-answer area. This enables the QA R-CNN to gain the ability to draw its attention to the area that the answer may appear in the image. As such, a straightforward idea is that the model can directly use the underlying text of the bounding box with the highest focusing score as the question’s answer. However, the rich semantics of the textual content will not be considered. Therefore, RM is introduced to further improve the pipeline. In RM, we follow the similar architecture in LoRRA where the semantics of detected text are further explored. Specifically, word embedding of the OCR tokens is extracted by FastText [66] models that are pre-trained on English/Chinese Wikipedia, and then the OCR embedding is fused with both image features and question embedding for further classification. Different from other classification-based approaches, we do not use a pre-defined fixed dictionary as the answer space but only use the detected OCR tokens, *i.e.*, only the detected text can be used as the answer. In the end, the weighted score of FM and RM are summed up for the final prediction.



FIGURE 5.10. Visualization of the output answers on the EST-VQA dataset from different models (first four images). Green and Red bounding boxes are ground-truth and predicted evidence by QA R-CNN. (More examples can be found at <https://arxiv.org/abs/2002.10215>)

5.3.2 Results

Quantitative Results: Table 5.3 summarizes the results of the baselines and our method on the EST-VQA dataset. The penalty threshold τ is practically set to 0.75 during the evaluation to ensure the answer quality. Figure 5.9 shows the CLC score under different τ for bilingual models.

We first measure the upper bound performance of the two pre-defined dictionaries SV and LV. Similar to other scene text VQA datasets, SV and LV can achieve high accuracy on English questions, *i.e.*, 31.1 and 48.0 respectively. However, they failed catastrophically on the Chinese questions due to the language features and lower overlapping of answers between the training and testing splits. Hence, it is more difficult for the classification based method to obtain a promising performance on the Chinese split in the EST-VQA dataset. We also provide the upper bound accuracy of the OCR results that are generated by [101, 143], and it achieves better accuracy on Chinese questions compared to the fixed vocabularies. Then a baseline using random OCR token is set as a comparison with other approaches, and this heuristic method only achieves 3.0 and 3.7 overall score for the CLC and TC tasks respectively.

To further justify the need for EST-VQA, we trained two state-of-the-art approaches, *i.e.*, Pythia (P) [150] and LoRRA (L) [151]. As shown in Table 5.3, both methods perform poorly on Chinese questions due to a large amount of out-of-vocabulary answers in the test set. Also, as the CLC task requires a model to provide evidence as well as the answer, the accuracy of all of the studied methods dropped significantly when compared to the TC score. This is because the models infer the answers without actually reading the textual content in the images (see Figure 5.3(c) and 5.3(d)), thus they can not provide reasonable evidence to support the answer. In contrast, the proposed QA R-CNN shows more robust results on the three tasks (see Table 5.3).

To further explore the proposed CLC task, we also trained a QA R-CNN with bells and whistles, many heuristic manual rules are adopted to lift the performance. Under this model, it outputs answers predominantly from the vocabulary for a certain type of questions. And if the model failed to detect the corresponding text, question related text would be picked up from the dictionary (e.g. digits for “what number”) as the answer. Although this heavy model achieves top performance on the TC task, its CLC score is even lower than the baseline QA R-CNN. Such a scenario suggests that the evaluation protocol used in the current conventional VQA task is not reasonable to some extent, because the VQA models can easily overfit to the answer space by using tricks. Therefore, we introduce a *reasonable score* Δ_r to measure the percentage of answers with sufficient evidence, and it is denoted as $\Delta_r = \frac{CLC_{all}}{TC_{all}}$. Lower Δ_r means that the model has outputted many unreasonable but correct answers, which suggests that it might either overfit to the answer pool or use too many manual rules to achieve a higher score under conventional evaluation protocol. As shown in Table 5.3, the QA R-CNN w/ tricks obtained the lowest reasonable score although it outperforms all other models under the traditional evaluation protocol. Another interesting observation is that all methods achieve extremely low accuracy on the questions that have a longer answer. We believe this is because current models cannot combine multiple texts to generate a long answer. However, how to solve this issue is out of the scope of this paper, and thus we leave it for future work.

Qualitative Results: Figure 5.10 illustrates some selected visualization results of the baseline methods. Surprisingly, we found that some models do not learn the concept of question type at all. For example, the ‘P+LV’ model outputs a word ‘caffe’ for the question ‘What is the room number?’ that asks for a number, and ‘L+LV’ predicts a character ‘长’ (long) for the question ‘这里是河南中路多少门牌号’ (What is the house number of this shop here in Henan Middle Road?) that is also asking a number. Furthermore, incorrect recognition results will cause the models to output incorrect answers. Based on the first sample of Figure 5.10, although the bounding box of the answer ‘708’ was predicted correctly, it was however recognized as ‘8’ and was further outputted as the answer. An interesting case is the ‘L+LV’ model answers the question ‘When was this photo uploaded?’ with ‘29/08/2012’ when only ‘2012’ appeared in the original image. Such a phenomenon tells us that similar answers in the vocabulary could interfere with the decision of classifier. Another noteworthy example is that ‘P+SV’ model predicts ‘snowbird’ for the question ‘When was this photo uploaded’. We queried another image with the answer ‘snowbird’

in the training set (see the last image in Figure 5.10) and it shows that the ‘P+SV’ model outputs the same answer when the image contains similar visual features. Therefore, we believe that this VQA model might rely too heavily on the image feature and learned to map the image feature with the answer space but it does not truly understand the question. Additionally, for the question that requires stronger reasoning ability and image with many texts, such as the third sample in Figure 5.10, ‘伟业水电安装的联系是谁? (Who is the contact person for Weiye Hydropower Installation?)’, none of the models are able to predict the answer correctly.

5.4 Conclusion

We have introduced a new bilingual scene text+evidence VQA dataset named EST-VQA that is annotated with both English and Chinese QA pairs. Three related challenges are proposed, namely *Cross Language*, *Localization* and *Traditional* that are designed to evaluate the generalization of VQA models. An evidence-based measure of an algorithm’s capacity to reason is also proposed that requires the VQA model to provide a bounding box for the predicted answer. This metric aims to uncover whether the VQA model learns deeper relationships between text and image content, rather than overfitting to a pre-defined dictionary. Future work includes extending the proposed EvE metric to existing VQA datasets in the hope that it might improve generalization and thus the practicality of VQA technologies.

5.5 Supplementary

5.5.1 Annotation Guidelines

Questions: Firstly, questions in STE-VQA were annotated according to the rules below:

1. All questions must be able to answer only by reading the textual contents in the images, for instance asking the license number of a specific car. Questions that violate this rule are prohibited (*e.g.* asking the colour of the vehicle).
2. Object in the question must be specified clearly so that it has an unambiguous answer. For example, generic questions such as what is the license number are not allowed as all vehicles have a license number. Instead, the

targeted vehicle should be pointed out through its colour or location and use the specific feature in question formulation.

3. The answer to the question must be able to retrieve from the image in textual format. It can be the texts on objects or even watermark on the image as long as the answer matches the textual content in the image.
4. Images with texts in multiple languages are allowed as long as the questions and answers can be annotated in English or Chinese language.
5. If any of the above question rules are not fulfilled, the image is discarded. Do note that, the preferred number of questions per image is 1 to 3, with a maximum of 5 questions.

Answers: Secondly, answers in STE-VQA are annotated adhere to the following rules:

1. The answers can be in the form of Latin characters, words, Arabic numerals, or any combination of them. Also, they are restricted to English or Chinese language only. Below is the list of acceptable answers:
 - Latin characters: a, b, A, z, P
 - English words: Phillips, Nike, Adidas, blackberry, Huawei
 - English phrases: the universal alamanac, near the town centre car parks
 - Arabic numerals: 44, 2019, 18, 16, 0
 - Combination: 13/1/2012, 00-b2w, o’neill, weston rd, bacanalnica.com, conan o’brien
2. Annotators are required to draw a rectangle or quadrilateral bounding box on the textual area of the image in which the bounding box must surround the texts tightly.

5.5.2 Annotation Pipeline

We show the annotation manual in Section 5.5.1, and two main stages of the data labelling progress in Figure 5.11 and Figure 5.13. Figure 5.11 shows the first stage of the pipeline which annotators are requested to label the bounding box for a potential question answer. Two modes of bounding-box are provided for labelling, *i.e.*, Rectangle and Quadrilateral. Annotators are asked to use the mode that can best fit the texts with the least blank space (see Figure 5.12

for example). Then in the second stage, for each image, the annotators are asked to come up with one to three questions that are related to the annotated text. Next, annotators are required to click on the bounding box listed on the right side of an image, and follow the format of ‘Q?A’ to input the question and answer pair. As illustrated in Figure 5.13, “What animal has been mentioned?Elephant”.

After filtering those images without text or appropriate question, Table 5.4 shows the final distribution of the number of questions in the STE-VQA dataset, which demonstrates that most images are annotated with a single question, this ensures the diversity of the proposed STE-VQA dataset.

		English				
# Q		1	2	3	4	5
		12,343	1,214	88	4	1
		Chinese				
# Q		1	2	3	4	5
		10,335	1,105	135	14	0

TABLE 5.4. Distribution of question number per image.

Figure 5.14 shows the word cloud of the majority answers in the STE-VQA dataset. We found that the answers which appear in the dataset most frequently are road signs, brands, places, numbers and etc., while most Chinese answers are locations, signboard names, numbers and etc. It is noteworthy that numbers and signboard names are of high variability, and can hardly be covered by the fixed vocabulary that is built from training set. Therefore, it would be extremely difficult for the traditional classification-based method to generalize to the test set. Besides, we show a more detailed composition of the answers in Table 5.5. It demonstrates that around 13% to 14% answers are pure numbers in English and Chinese questions, and 34% of English answers are comprised of more than two words while about 40% Chinese answers are of long phrases. A notable fact is shorter numbers such as ‘1’, ‘2’, ‘4’ repeat more frequently, thus they can be covered easily in the pre-generated dictionary. However, longer numbers such as mobile phone numbers and series numbers are less likely to occur in training and testing set at the same time. Specifically, 2617 out of 3895 pure-digit answers have more than 3 numbers in the whole STE-VQA dataset, which are extremely challenging for both the OCR and QA parts of text-based VQA models.

English				
# A	String	Num	Short	Long
	12,840	2,216	9,899	5,157
Chinese				
# A	String	Num	Short	Long
	11,327	1,679	8,896	6,160

TABLE 5.5. Distribution of answer type in the STE-VQA dataset.



FIGURE 5.11. Labelling Tool. At the first stage, annotators are asked to label a rectangle or quadrilateral bounding-box for a potential answer.

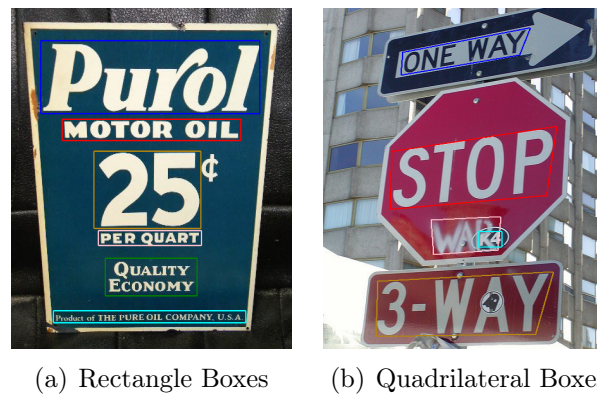


FIGURE 5.12. Two bounding-box labelling modes are available in the annotation tool. Annotators are asked to select the most appropriate one by considering the tightness between the text and bounding box.

5.6 More Annotation Examples

Figure 5.15 and Figure 5.16 show more examples of English questions and Chinese questions, respectively. Scenes with English or Chinese texts share



Q: Where is here?

A: Summer Place

(a)



Q: Where are vehical not allowed to enter?

A: Campus

(b)



Q: What is the name of this project?

A: GREAT WALL PLAN

(c)



Q: What's the phone number of this store?

A: 626-8727

(d)

FIGURE 5.15. Examples of English questions.



Q: 这里是什么地方?

A: 新源一村

(a)



Q: 最左边的门牌写着什么?

A: 北京市癫痫病诊疗中心

(b)



Q: 左边的店的名字是什么?

A: 郑精五金机电

(c)



Q: 卫青汽修的电话号码是什么?

A: 17682439669

(d)



Q: 这个小区的名字是什么?

A: 自由四村

(e)



Q: 北京翔达投资管理有限公司的地址在哪里?

A: 北京市宣武区教子胡同 28 号

(f)



Q: 这里是哪本杂志的编辑部?

A: 骨科临床与研究杂志

(g)



Q: 这栋建筑的名字是什么?

A: 隔山祖厝

(h)

FIGURE 5.16. Examples of Chinese questions.

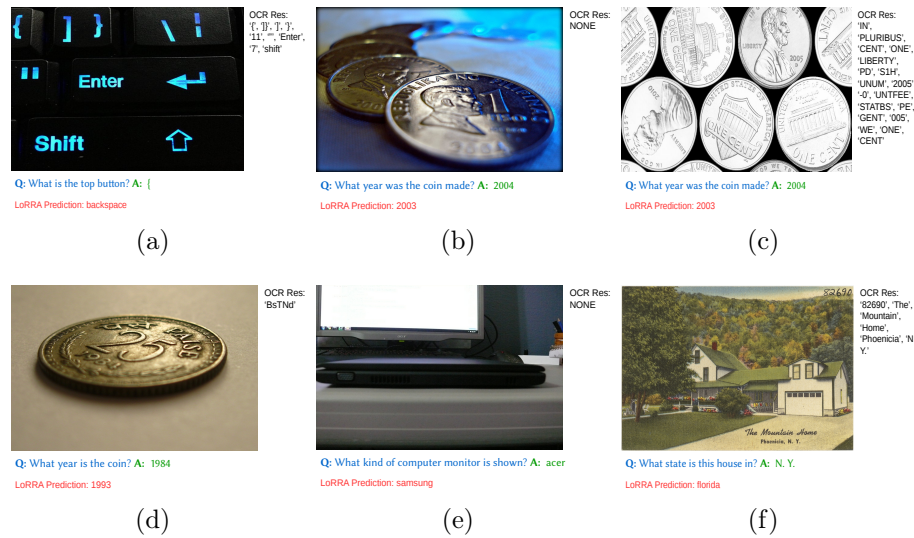


FIGURE 5.17. Unreasonable Output Part A

the question ‘这栋建筑的名字是什么(What is the name of this building?)’ is ‘隔山祖庭’ but not ‘庭祖山隔’. This example reflects the grammatical and cultural distinctions between English and Chinese languages, which further encourages the generalization ability of the VQA models. Based on this, a new research question is raised with the introduction of STE-VQA, *i.e.*, how to design a content-aware OCR system that can decide the textual order for the contents written in traditional style.

5.7 More Examples of Unreasonable Output in Conventional Text-VQA dataset

In Figure 3 of the original paper, we show some examples of unreasonable predictions outputted by the traditional none-evidence-based methods LoRRA [151]. It was trained on the Text-VQA dataset [151], which suggests that the conventional VQA approaches are prone to learn coincidental correlations in the data. Here we show more examples in Figure 5.17 and Figure 5.18. Figure 5.17(b), 5.17(c), and 5.17(d) show a similar situation, the questions are all about the ‘year’ when was the coin made, and the OCR system failed to recognize the correct answer in these three images. Although LoRRA failed to answer these questions correctly, we could find that the predicted answers still have a strong relation to the question, *i.e.*, all of the predictions are actually ‘year’. A similar scenario happened in Figure 5.17(a), the question asks for a keyboard button and the model predicted a keyboard button ‘backspace’ that



FIGURE 5.18. Unreasonable Output Part B

does not appear in the image, though the correct answer ‘{’ was detected by the OCR system. In Figure 5.17(e), the question asks for a brand of a computer monitor and the model outputs a brand ‘samsung’ that does not appear in the image; and in Figure 5.17(f), the question asks for a state and the model outputs a state ‘Florida’ that does not appear in the image. Also, in Figure 5.18 we show that the model could answer the question correctly without correct OCR results. All of the samples in Figure 5.18 are answered correctly without correct OCR input. Specifically, in Figure 5.18(a), it can be noticed that the brand of the soda is occluded, but the LoRRA model is still able to predict a correct answer. Two similar questions are shared by the samples shown in Figure 5.18(d) and 5.18(e), which ask for the number located at the top left corner. The model outputs the correct answer ‘1’ for both images even though the OCR system failed to detect this number. Based on these observations, it suggests that traditional approaches focus on the coincidental correlations between features and answers but not truly learn to reason. For example, the model might learn the visual feature of the can in Figure 5.18(a) to answer the question, but it does not read the brand; while the model might learn the possible relation between the word ‘queen’ in the question shown in Figure 5.18(c) and choose ‘Elizabeth’ to answer the question. Therefore, it is necessary to use the EvE evaluation protocol that is proposed in STE-VQA, because result-oriented methods might not be able to measure the actual VQA models’ capability to understand and reason about questions.

Finally, we summarize and highlight the differences in several aspects of the aforementioned datasets in Table 5.6.

Aspect	TextVQA [151]	ST-VQA [13]	STE-VQA (ours)
Image source(s)	Images are collected from Open Images V3 [76] dataset only which the text instances have limited orientations as the primary focus of images is not on scene text.	Images are from a combination of image datasets with multiple objectives such as image classification, VQA, scene text, etc. Text orientations available in this dataset might be better than TextVQA.	Images are all from public scene text datasets that contain a variety of text orientations. This poses a harder challenge for the models as they are required to deal with scene text images that can be commonly found in real life.
Question and answer pairs	Open-ended questions are asked in this dataset that accepts inferred or paraphrased answers based on textual contents in the images. Binary answer (<i>e.g.</i> yes/no) is acceptable and takes up of 5.55% of the entire dataset.	Annotators are encouraged to ask close-ended questions where their answers must be found as texts in the images, yes/no questions are prohibited.	Close-ended questions are asked only in which the answers must be able to read from the images. Evidence or bounding box of the text instances are provided as well apart from the usual question and answer pairs.
Language(s)		English only.	English and Chinese. Introduces the text reading sequence problem of scene text detection and recognition to VQA (<i>e.g.</i> left to right or right to left) which requires the VQA model to understand the semantic context of the answers based on different languages.
Tasks	Does not introduce any new tasks.	Introduces three tasks with varying vocabulary sizes.	Three tasks are introduced that aim to evaluate different aspects of the model, such as the capability to deal with multiple languages and the ability to provide supporting evidence.
Evaluation metric	Employs the commonly used VQA accuracy as introduced in [43]. It does not take in the recognition performance of the OCR module.	Proposed Average Normalized Levenshtein Similarity (ANLS) as the evaluation metric used for all 3 tasks. It is commonly used in scene text recognition that penalizes the output based on the normalized edit distance from the ground truth. However, it does not consider how the model reason about the output through additional measures.	Evidence-based Evaluation (EvE) metric is proposed in which the model will be evaluated first based on the outputted answer (using ANLS) and then the bounding box used to support its answer (using Intersection over Union). The final score will only consider outputted answer with sufficient bounding box as evidence only. This increases the difficulty of STE-VQA as models are required to be able to provide bounding box along with answer simultaneously.

TABLE 5.6. A summary and comparison of different aspects between Text-VQA, ST-VQA and the proposed STE-VQA.

Chapter 6

Conclusions

Benefiting from the rapid development of deep learning techniques, OCR technologies have witnessed significant advancements in recent years. However, although promising performance can be achieved in a variety of datasets, OCR systems still face challenges when dealing with complex and diverse text images, such as low-quality images, handwriting, or non-Latin scripts. In this thesis, we reviewed several existing issues in the OCR field and explored potential solutions through three aspects, i.e., 1) designing a unified benchmarking framework to enable fair comparisons between different OCR approaches; 2) developing a novel text synthesis method to mitigate the issue of imbalanced data; 3) proposing an evidence-based framework for text visual question answering model, boosting the model’s reasoning capability.

First, we reviewed a number of recently proposed OCR papers and found that unfair comparisons were often made between different models. For example, some papers would use a specific dataset or set of parameters that favored their own model, while others would use a different set of training and testing parameters. This made it difficult to accurately compare the performance of different OCR models. To address this issue, we proposed a standardized evaluation framework called the UniOCR benchmark. This framework was designed from three aspects, i.e., datasets, metrics, and models, to ensure the fairness and consistency of OCR model comparisons. Specifically, the UniOCR benchmark includes a diverse set of datasets covering different languages and annotation forms. This allows for a comprehensive evaluation of OCR models across a range of different scenarios. Additionally, the benchmark includes a set of standardized metrics, enabling more accurate measurement of texts with different lengths. Furthermore, a standard pipeline for training and testing stages was established, ensuring that all models are evaluated under the same conditions and with the same parameters. By providing standardized evaluation suites, the UniOCR benchmark aims to promote fair and consistent comparisons between OCR models and to facilitate the development and improvement

of future OCR-related research.

Moreover, we propose a text-to-image synthesis framework called TLPNet to address the unbalanced data issue in the driving license plate recognition task. A non-negligible issue that exists in this task is that the distribution of data samples is highly correlated with the location where the images were collected. For example, if the majority of the images were collected in a specific city or region, the license plate data would be heavily skewed toward that area’s license plate styles, characters, and region codes. This can lead to poor recognition performance for license plates from other areas. To tackle this problem, TLPNet develops a text-to-image network to synthesize photo-realistic license plate samples, enabling a more balanced and diverse training dataset. This allows the model to better handle a wide range of license plate styles and characters, improving overall recognition performance.

Finally, in addition to identifying texts within images, it is crucial for models to comprehend the meaning behind the text. This poses a significant obstacle as it necessitates a model to possess a deep understanding of both visual and textual information. A persistent problem in this text-based visual question answering is the lack of reasoning ability in current models. This means that VQA models simply predict answers without offering any reasoning or justification, leading to poor generalizability. To overcome this challenge, we propose to explore evidence-based text visual question answering, which entails designing models that can provide reasoning and evidence for their predictions, thereby enhancing their generalization ability and robustness to unseen examples. Additionally, we introduce a new dataset and a novel metric to quantitatively evaluate a model’s reasoning capability.

We are of the opinion that the techniques outlined in this thesis have the potential to alter the current state of optical character recognition. And it is our hope that our proposed datasets, methods, and benchmarks can serve as a sturdy foundation for various endeavors and uses that necessitate text detection and recognition and provide a fresh perspective for the community.

6.1 Future Work

In light of the aforementioned findings and contributions, the next step for future research work would be to explore the potential synergies between OCR models and language models. Considering the current landscape of OCR research, it is worth noting that most existing methods primarily rely on visual information for text recognition. However, text, in its very essence, contains rich

semantic information, and human reading behavior naturally involves leveraging semantic context to aid in the recognition or completion of the ambiguous or damaged text. This leads us to contemplate the potential benefits of incorporating large-scale pre-trained language models, which inherently possess powerful semantic understanding into the OCR domain. Therefore, a promising research direction lies in the integration of visual information into large pre-trained language models to assist in training OCR models. This approach could not only enhance the recognition capabilities of OCR systems by tapping into the rich semantic information present in language models but also enable few-shot learning techniques to adapt multi-modal pre-trained models for OCR tasks. In doing so, we would harness the inherent strengths of language models to create more robust and adaptable OCR systems that can effectively handle diverse and complex text images. Thus, by incorporating the knowledge of large language models and leveraging few-shot learning for multi-modal model adaptation, we hope to break new ground in OCR research, ultimately leading to more accurate and versatile systems capable of understanding and interpreting text across a wide range of scenarios.

Bibliography

- [1] <https://github.com/derek285/generateCarPlate>.
- [2] <https://github.com/Belval/TextRecognitionDataGenerator>, 2022.
- [3] <https://github.com/ultralytics/yolov5>, 2022.
- [4] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4971–4980, 2018.
- [5] Christos Nikolaos E Anagnostopoulos, Ioannis E Anagnostopoulos, Vassilis Loumos, and Eleftherios Kayafas. A license plate-recognition algorithm for intelligent transportation system applications. *IEEE Trans. Intell. Transp. Syst.*, 7(3):377–392, 2006.
- [6] Christos-Nikolaos E Anagnostopoulos, Ioannis E Anagnostopoulos, Ioannis D Psoroulas, Vassili Loumos, and Eleftherios Kayafas. License plate recognition from still images and video sequences: A survey. *IEEE Trans. Intell. Transp. Syst.*, 9(3):377–391, 2008.
- [7] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3674–3683, 2018.
- [8] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 2425–2433, 2015.
- [9] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2019.

-
- [10] Jeonghun Baek, Yusuke Matsui, and Kiyoharu Aizawa. What if we only use real datasets for scene text recognition? toward scene text recognition with fewer labels. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3113–3122, 2021.
- [11] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoon Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 9365–9374, 2019.
- [12] Fan Bai, Zhazhan Cheng, Yi Niu, Shiliang Pu, and Shuigeng Zhou. Edit probability for scene text recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1508–1516, 2018.
- [13] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluís Gomez, Marçal Rusiñol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. *Proc. IEEE Int. Conf. Comp. Vis.*, 2019.
- [14] Tomas Björklund, Attilio Fiandrotti, Mauro Annarumma, Gianluca Francini, and Enrico Magli. Robust license plate recognition using neural networks trained on synthetic images. *Pattern Recogn.*, 93:134–146, 2019.
- [15] Syed Saqib Bukhari, Ahmad Kadi, Mohammad Ayman Jouneh, Fahim Mahmood Mir, and Andreas Dengel. anyocr: An open-source ocr system for historical archives. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, volume 1, pages 305–310. IEEE, 2017.
- [16] Orhan Bulan, Vladimir Kozitsky, Palghat Ramesh, and Matthew Shreve. Segmentation-and annotation-free license plate recognition with deep localization and failure identification. *IEEE Trans. Intell. Transp. Syst.*, 18(9):2351–2363, 2017.
- [17] Jean-Christophe Burie, Joseph Chazalon, Mickaël Coustaty, Sébastien Eskenazi, Muhammad Muzzamil Luqman, Maroua Mehri, Nibal Nayef, Jean-Marc Ogier, Sophea Prum, and Marçal Rusiñol. Icdar2015 competition on smartphone document capture and ocr (smartdoc). In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 1161–1165. IEEE, 2015.
- [18] Shyang-Lih Chang, Li-Shien Chen, Yun-Chung Chung, and Sei-Wan Chen. Automatic license plate recognition. *IEEE Trans. Intell. Transp. Syst.*, 5(1):42–53, 2004.
- [19] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi

- Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.
- [20] Xiaoxue Chen, Lianwen Jin, Yuanzhi Zhu, Canjie Luo, and Tianwei Wang. Text recognition in the wild: A survey. *CSUR*, 54(2):1–35, 2021.
- [21] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 5076–5084, 2017.
- [22] Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, and Shuigeng Zhou. AON: Towards arbitrarily-oriented text recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5571–5579, 2018.
- [23] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.
- [24] Chee Kheng Ch’ng and Chee Seng Chan. Total-Text: A comprehensive dataset for scene text detection and recognition. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 935–942, 2017.
- [25] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. ICDAR2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 1571–1576, 2019.
- [26] Chee Kheng Ch’ng, Chee Seng Chan, and Chenglin Liu. Total-text: Towards orientation robustness in scene text detection. *Int. J. Doc. Anal. Recognit.*, 23:31–52, 2020. doi: 10.1007/s10032-019-00334-z.
- [27] Yuchen Dai, Zheng Huang, Yuting Gao, Youxuan Xu, Kai Chen, Jie Guo, and Weidong Qiu. Fused text segmentation networks for multi-oriented scene text detection. In *Proc. Int. Conf. Patt. Recogn.*, pages 3604–3609. IEEE, 2018.
- [28] Brian Davis, Bryan Morse, Scott Cohen, Brian Price, and Chris Tensmeyer. Deep visual template-free form parsing. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 134–141, 2019.

- [29] Alysson de Sá Soares, Ricardo Batista das Neves Junior, and Byron Leite Dantas Bezerra. BID dataset: a challenge dataset for document processing tasks. In *GPI*, pages 143–146, 2020.
- [30] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 248–255. Ieee, 2009.
- [31] Shan Du, Mahmoud Ibrahim, Mohamed Shehata, and Wael Badawy. Automatic license plate recognition (alpr): A state-of-the-art review. *IEEE Trans. Circuits Syst. Video Technol.*, 23(2):311–325, 2012.
- [32] Kartik Dutta, Minesh Mathew, Praveen Krishnan, and CV Jawahar. Localizing and recognizing text in lecture videos. In *Proc. Int. Conf. Front. Handwrit. Recognit.*, pages 235–240, 2018.
- [33] Salvador Espana-Boquera, Maria Jose Castro-Bleda, Jorge Gorbe-Moya, and Francisco Zamora-Martinez. Improving offline handwritten text recognition with hybrid hmm/ann models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(4):767–779, 2010.
- [34] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.*, 111(1):98–136, 2015.
- [35] Ross Girshick. Fast r-cnn. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1440–1448, 2015.
- [36] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 580–587, 2014.
- [37] Lluís Gómez, Andrés Mafla, Marçal Rusinol, and Dimosthenis Karatzas. Single shot scene text retrieval. In *Proc. Eur. Conf. Comp. Vis.*, pages 700–715, 2018.
- [38] Raul Gomez, Baoguang Shi, Lluís Gomez, Lukas Numann, Andreas Veit, Jiri Matas, Serge Belongie, and Dimosthenis Karatzas. ICDAR2017 robust reading challenge on coco-text. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 1435–1443, 2017.

- [39] Gabriel Resende Gonçalves, Sirlene Pio Gomes da Silva, David Menotti, and William Robson Schwartz. Benchmark for license plate character segmentation. *Journal of Electronic Imaging*, 25(5):053034, 2016.
- [40] Santhoshini Gongidi and CV Jawahar. IIIT-indic-hw-words: A dataset for indic handwritten text recognition. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 444–459, 2021.
- [41] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Proc. Advances in Neural Inf. Process. Syst.*, 2014.
- [42] Chao Gou, Kunfeng Wang, Yanjie Yao, and Zhengxi Li. Vehicle license plate recognition based on extremal regions and restricted boltzmann machines. *IEEE Trans. Intell. Transp. Syst.*, 17(4):1096–1107, 2015.
- [43] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 6904–6913, 2017.
- [44] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. Int. Conf. Mach. Learn.*, pages 369–376. ACM, 2006.
- [45] Alex Graves, Marcus Liwicki, Horst Bunke, Jürgen Schmidhuber, and Santiago Fernández. Unconstrained on-line handwriting recognition with recurrent neural networks. *Proc. Advances in Neural Inf. Process. Syst.*, 20, 2007.
- [46] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Proc. IEEE Int. Conf. Acoustics, Speech, & Signal Process.*, pages 6645–6649. Ieee, 2013.
- [47] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *Proc. Int. Conf. Mach. Learn.*, pages 1462–1471. PMLR, 2015.

- [48] Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. Unidoc: Unified pre-training framework for document understanding. *Proc. Advances in Neural Inf. Process. Syst.*, 34:39–50, 2021.
- [49] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2315–2324, 2016.
- [50] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3608–3617, 2018.
- [51] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 770–778, 2016.
- [52] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 2961–2969, 2017.
- [53] Mengchao He, Yuliang Liu, Zhibo Yang, Sheng Zhang, Canjie Luo, Feiyu Gao, Qi Zheng, Yongpan Wang, Xin Zhang, and Lianwen Jin. ICPR2018 contest on robust reading for multi-type web images. In *Proc. Int. Conf. Patt. Recogn.*, pages 7–12, 2018.
- [54] Pan He, Weilin Huang, Tong He, Qile Zhu, Yu Qiao, and Xiaolin Li. Single shot text detector with regional attention. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3047–3055, 2017.
- [55] Gee-Sern Hsu, Jiun-Chang Chen, and Yu-Zu Chung. Application-oriented license plate recognition. *IEEE Trans. Veh. Technol.*, 2012.
- [56] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 1516–1520. IEEE, 2019.
- [57] Masakazu Iwamura, Takahiro Matsuda, Naoyuki Morimoto, Hitomi Sato, Yuki Ikeda, and Koichi Kise. Downtown osaka scene text dataset. In *Proc. Eur. Conf. Comp. Vis.*, pages 440–455, 2016.

- [58] Brian Kenji Iwana, Syed Tahseen Raza Rizvi, Sheraz Ahmed, Andreas Dengel, and Seiichi Uchida. Judging a book by its cover. *arXiv preprint arXiv:1610.09204*, 2016.
- [59] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014.
- [60] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Proc. Advances in Neural Inf. Process. Syst.*, 28:2017–2025, 2015.
- [61] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *Int. J. Comput. Vis.*, 116(1):1–20, 2016.
- [62] Soumya Jahagirdar, Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Watching the news: Towards videoqa models that can read. In *Proc. Winter Conf. Appl. of Comp. Vis.*, pages 4441–4450, 2023.
- [63] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. FUNSD: A dataset for form understanding in noisy scanned documents. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 1–6, 2019.
- [64] Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu, and Zhenbo Luo. R2cnn: Rotational region cnn for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579*, 2017.
- [65] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.
- [66] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *European Chapter of the Association for Computational Linguistics*, 2016.
- [67] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Comput. Vis. Image Underst.*, 163:3–20, 2017.

- [68] Sezer Karaoglu, Ran Tao, Theo Gevers, and Arnold WM Smeulders. Words matter: Scene text for image classification and retrieval. *IEEE Trans. Multimedia*, 19(5):1063–1076, 2016.
- [69] Dimosthenis Karatzas, S Robles Mestre, Joan Mas, Farshad Nourbakhsh, and P Pratim Roy. ICDAR 2011 robust reading competition-challenge 1: reading text in born-digital images (web and email). In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 1485–1490, 2011.
- [70] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluís Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluís Pere De Las Heras. Icdar 2013 robust reading competition. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 1484–1493. IEEE, 2013.
- [71] Dimosthenis Karatzas, Lluís Gomez-Bigorda, Angelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 1156–1160. IEEE, 2015.
- [72] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4999–5007, 2017.
- [73] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proc. Int. Conf. Learn. Representations*, 2015.
- [74] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [75] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *Proc. Int. Conf. Learn. Representations*, 2020.
- [76] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from <https://github.com/openimages>*, 2017.

- [77] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017.
- [78] Praveen Krishnan, Rama Kovvuri, Guan Pang, Boris Vassilev, and Tal Hassner. Textstylebrush: transfer of text aesthetics from a single example. *arXiv preprint arXiv:2106.08385*, 2021.
- [79] Alex W. C. Lee, Jonathan Chung, and Marco Lee. GNHK: A dataset for english handwriting in the wild. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, 2021.
- [80] Chen-Yu Lee and Simon Osindero. Recursive recurrent nets with attention modeling for ocr in the wild. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2231–2239, 2016.
- [81] SeongHun Lee, Min Su Cho, Kyomin Jung, and Jin Hyung Kim. Scene text extraction with edge constraint and text collinearity. In *Proc. Int. Conf. Patt. Recogn.*, pages 3983–3986, 2010.
- [82] Vladimir Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.
- [83] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Controllable text-to-image generation. *Proc. Advances in Neural Inf. Process. Syst.*, 2019.
- [84] Hui Li and Chunhua Shen. Reading car license plates using deep convolutional neural networks and lstms. *arXiv preprint arXiv:1601.05610*, 2016.
- [85] Hui Li, Peng Wang, and Chunhua Shen. Towards end-to-end text spotting with convolutional recurrent neural networks. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 5238–5246, 2017.
- [86] Hui Li, Peng Wang, and Chunhua Shen. Toward end-to-end car license plate detection and recognition with deep neural networks. *IEEE Trans. Intell. Transp. Syst.*, 20(3):1126–1136, 2018.

- [87] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. In *Proc. AAAI Conf. Artificial Intell.*, pages 8610–8617, 2019.
- [88] Piyuan Li, Minh Nguyen, and Wei Qi Yan. Rotation correction for license plate recognition. In *Proc. Int. Conf. Control, Automation and Robotics*, pages 400–404, 2018.
- [89] Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. Analogical reasoning on chinese morphological and semantic relations. *Proc. Annu. Meet. Assoc. Comput. Linguist.*, 2018.
- [90] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *Proc. AAAI Conf. Artificial Intell.*, 2017.
- [91] Minghui Liao, Guan Pang, Jing Huang, Tal Hassner, and Xiang Bai. Mask TextSpotter v3: Segmentation proposal network for robust scene text spotting. In *Proc. Eur. Conf. Comp. Vis.*, pages 706–722, 2020.
- [92] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *Proc. AAAI Conf. Artificial Intell.*, pages 11474–11481, 2020.
- [93] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. Eur. Conf. Comp. Vis.*, pages 740–755, 2014.
- [94] Cheng-Lin Liu, Fei Yin, Da-Han Wang, and Qiu-Feng Wang. CASIA online and offline chinese handwriting databases. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 37–41, 2011.
- [95] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *Proc. Eur. Conf. Comp. Vis.*, pages 21–37. Springer, 2016.
- [96] Wei Liu, Chaofeng Chen, Kwan-Yee K Wong, Zhizhong Su, and Junyu Han. Star-net: a spatial attention residue network for scene text recognition. In *Proc. British Machine Vis. Conf.*, volume 2, page 7, 2016.
- [97] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5676–5685, 2018.

-
- [98] Yuliang Liu and Lianwen Jin. Deep matching prior network: Toward tighter multi-oriented text detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.
- [99] Yuliang Liu, Lianwen Jin, Zecheng Xie, Canjie Luo, Shuaitao Zhang, and Lele Xie. Tightness-aware evaluation protocol for scene text detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 9612–9620, 2019.
- [100] Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Sheng Zhang. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recogn.*, 90:337–345, 2019.
- [101] Yuliang Liu, Sheng Zhang, Lianwen Jin, Lele Xie, Yaqiang Wu, and Zhepeng Wang. Omnidirectional scene text detection with sequential-free box discretization. *Proc. Int. Joint Conf. Artificial Intell.*, 2019.
- [102] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 9809–9818, 2020.
- [103] Yuliang Liu, Tong He, Hao Chen, Xinyu Wang, Canjie Luo, Shuaitao Zhang, Chunhua Shen, and Lianwen Jin. Exploring the capacity of an orderless box discretization network for multi-orientation scene text detection. *Int. J. Comput. Vis.*, 129(6):1972–1992, 2021.
- [104] Yuliang Liu, Jiaxin Zhang, Dezhi Peng, Mingxin Huang, Xinyu Wang, Jingqun Tang, Can Huang, Dahua Lin, Chunhua Shen, Xiang Bai, et al. Spts v2: Single-point scene text spotting. *arXiv preprint arXiv:2301.01635*, 2023.
- [105] Zichuan Liu, Yixing Li, Fengbo Ren, Wang Ling Goh, and Hao Yu. Squeezedtext: A real-time scene text recognition by binary convolutional encoder-decoder network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [106] Shangbang Long, Xin He, and Cong Yao. Scene text detection and recognition: The deep learning era. *Int. J. Comput. Vis.*, 129(1):161–184, 2021.
- [107] Simon M Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, et al. Icdar 2003 robust reading competitions: entries, results, and future directions. *Int. J. Doc. Anal. Recognit.*, 7(2):105–122, 2005.

- [108] Canjie Luo, Lianwen Jin, and Zenghui Sun. Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recogn.*, 90: 109–118, 2019.
- [109] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *Proc. Eur. Conf. Comp. Vis.*, pages 67–83, 2018.
- [110] Pengyuan Lyu, Cong Yao, Wenhao Wu, Shuicheng Yan, and Xiang Bai. Multi-oriented scene text detection via corner localization and region segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 7553–7563, 2018.
- [111] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. Multimedia*, 20(11):3111–3122, 2018.
- [112] Andrés Mafla, Rafael S Rezende, Lluís Gomez, Diane Larlus, and Dimosthenis Karatzas. Stacmr: Scene-text aware cross-modal retrieval. In *Proc. Winter Conf. Appl. of Comp. Vis.*, pages 2220–2230, 2021.
- [113] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 1682–1690, 2014.
- [114] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *Proc. Int. Conf. Learn. Representations*, 2016.
- [115] Minesh Mathew, Mohit Jain, and CV Jawahar. Benchmarking scene text recognition in devanagari, telugu and malayalam. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, volume 7, pages 42–46, 2017.
- [116] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proc. Winter Conf. Appl. of Comp. Vis.*, pages 2200–2209, 2021.
- [117] Anand Mishra, Kartteek Alahari, and CV Jawahar. Image retrieval using textual cues. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 3040–3047, 2013.

- [118] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 947–952. IEEE, 2019.
- [119] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, 2019.
- [120] Harold Mouchère, Christian Viard-Gaudin, Richard Zanibbi, and Utpal Garain. Icfhr2016 crohme: Competition on recognition of online handwritten mathematical expressions. In *Proc. Int. Conf. Front. Handwrit. Recognit.*, pages 607–612. IEEE, 2016.
- [121] Robert Nagy, Anders Dicker, and Klaus Meyer-Wegener. Neocr: A configurable dataset for natural image text recognition. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 150–163. Springer, 2012.
- [122] Toshiaki Nakamura, Anna Zhu, Keiji Yanai, and Seiichi Uchida. Scene text eraser. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, volume 1, pages 832–837. IEEE, 2017.
- [123] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, volume 1, pages 1454–1459, 2017.
- [124] Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khelif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-lin Liu, et al. ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 1582–1587, 2019.
- [125] Lukáš Neumann and Jiří Matas. Real-time scene text localization and recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3538–3545. IEEE, 2012.
- [126] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4467–4477, 2017.

- [127] Nguyen Nguyen, Thu Nguyen, Vinh Tran, Minh-Triet Tran, Thanh Duc Ngo, Thien Huu Nguyen, and Minh Hoai. Dictionary-guided scene text recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 7383–7392, 2021.
- [128] Mohammad Tanvir Parvez and Sabri A Mahmoud. Offline arabic handwritten text recognition: a survey. *ACM Comput. Surveys*, 45(2):1–35, 2013.
- [129] Dezhi Peng, Xinyu Wang, Yuliang Liu, Jiaxin Zhang, Mingxin Huang, Songxuan Lai, Jing Li, Shenggao Zhu, Dahua Lin, Chunhua Shen, et al. Spts: single-point text spotting. In *Proc. ACM Int. Conf. Multimedia*, pages 4272–4281, 2022.
- [130] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proc. Conf. Empir. Methods in Natural Language Process.*, pages 1532–1543, 2014.
- [131] Ioannis Pratikakis, Konstantinos Zagoris, Basilis Gatos, Joan Puigcerver, Alejandro H Toselli, and Enrique Vidal. ICFHR2016 handwritten keyword spotting competition (h-kws 2016). In *Proc. Int. Conf. Front. Handwrit. Recognit.*, pages 613–618, 2016.
- [132] Joan Puigcerver. Are multidimensional recurrent layers really necessary for handwritten text recognition? In *Proc. Int. Conf. Doc. Anal. and Recognit.*, volume 1, pages 67–72. IEEE, 2017.
- [133] Liang Qiao, Sanli Tang, Zhazhan Cheng, Yunlu Xu, Yi Niu, Shiliang Pu, and Fei Wu. Text perceptron: Towards end-to-end arbitrary-shaped text spotting. In *Proc. AAAI Conf. Artificial Intell.*, pages 11899–11907, 2020.
- [134] Liang Qiao, Ying Chen, Zhazhan Cheng, Yunlu Xu, Yi Niu, Shiliang Pu, and Fei Wu. MANGO: A mask attention guided one-stage scene text spotter. *Proc. AAAI Conf. Artificial Intell.*, 2021.
- [135] Siyang Qin, Alessandro Bissacco, Michalis Raptis, Yasuhisa Fujii, and Ying Xiao. Towards unconstrained end-to-end text spotting. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 4704–4714, 2019.
- [136] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin,

- Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.
- [137] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.
- [138] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 14866–14876, 2019.
- [139] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 7263–7271, 2017.
- [140] Scott Reed, Zeynep Akata, Xinchun Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proc. Int. Conf. Mach. Learn.*, pages 1060–1069. PMLR, 2016.
- [141] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 91–99, 2015.
- [142] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *ESWA*, 41(18):8027–8048, 2014.
- [143] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2298–2304, 2016.
- [144] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4168–4176, 2016.
- [145] Baoguang Shi, Xiang Bai, and Serge Belongie. Detecting oriented text in natural images by linking segments. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.
- [146] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2298–2304, 2017.

- [147] Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. ICDAR2017 competition on reading chinese text in the wild (rctw-17). In *Proc. Int. Conf. Doc. Anal. and Recognit.*, volume 1, pages 1429–1434, 2017.
- [148] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.
- [149] Sérgio Montazzolli Silva and Cláudio Rosito Jung. License plate detection and recognition in unconstrained scenarios. In *Proc. Eur. Conf. Comp. Vis.*, pages 580–596, 2018.
- [150] Amanpreet Singh, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia-a platform for vision & language research. In *SysML Workshop, Advances in Neural Inf. Process. Syst.*, 2018.
- [151] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 8317–8326, 2019.
- [152] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 8802–8812, 2021.
- [153] Bolan Su and Shijian Lu. Accurate scene text recognition based on recurrent neural network. In *Proc. Asian Conf. Comp. Vis.*, pages 35–48. Springer, 2015.
- [154] Bolan Su and Shijian Lu. Accurate recognition of words in scenes without character segmentation using recurrent neural network. *Pattern Recogn.*, 63:397–405, 2017.
- [155] Junyi Sun. ‘Jieba’. <https://github.com/fxsjy/jieba>, 2012.
- [156] Yipeng Sun, Jiaming Liu, Wei Liu, Junyu Han, Errui Ding, and Jingtuo Liu. Chinese street view text: Large-scale chinese text reading with partially supervised learning. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 9086–9095, 2019.

- [157] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Proc. Advances in Neural Inf. Process. Syst.*, 27, 2014.
- [158] Jun Tang, Zhibo Yang, Yongpan Wang, Qi Zheng, Yongchao Xu, and Xiang Bai. Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping. *Pattern Recogn.*, 96:106954, 2019.
- [159] Gustav Tauschek. Reading machine, Apr 1938.
- [160] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. Detecting text in natural image with connectionist text proposal network. In *Proc. Eur. Conf. Comp. Vis.*, pages 56–72, 2016.
- [161] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.
- [162] Fangfang Wang, Liming Zhao, Xi Li, Xinchao Wang, and Dacheng Tao. Geometry-aware scene text detection with instance transformation network. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1381–1389, 2018.
- [163] Fangfang Wang, Yifeng Chen, Fei Wu, and Xi Li. Textray: Contour-based geometric modeling for arbitrary-shaped scene text detection. In *Proc. ACM Int. Conf. Multimedia*, pages 111–119, 2020.
- [164] Hao Wang, Xiang Bai, Mingkun Yang, Shenggao Zhu, Jing Wang, and Wenyu Liu. Scene text retrieval via joint text detection and similarity learning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4558–4567, 2021.
- [165] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1457–1464. IEEE, 2011.
- [166] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. FVQA: Fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10):2413–2427, October 2018. ISSN 0162-8828. doi: 10.1109/TPAMI.2017.2754246. URL <https://doi.org/10.1109/TPAMI.2017.2754246>.

- [167] Pengfei Wang, Chengquan Zhang, Fei Qi, Shanshan Liu, Xiaoqiang Zhang, Pengyuan Lyu, Junyu Han, Jingtuo Liu, Errui Ding, and Guangming Shi. PGNet: Real-time arbitrarily-shaped text spotting with point gathering network. In *Proc. AAAI Conf. Artificial Intell.*, 2021.
- [168] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 8198–8207, 2019.
- [169] Tao Wang, David J Wu, Adam Coates, and Andrew Y Ng. End-to-end text recognition with convolutional neural networks. In *Proc. Int. Conf. Patt. Recogn.*, pages 3304–3308. IEEE, 2012.
- [170] Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, and Mingxiang Cai. Decoupled attention network for text recognition. In *Proc. AAAI Conf. Artificial Intell.*, pages 12216–12224, 2020.
- [171] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 9336–9345, 2019.
- [172] Wenhai Wang, Enze Xie, Xiang Li, Xuebo Liu, Ding Liang, Yang Zhibo, Tong Lu, and Chunhua Shen. PAN++: Towards efficient and accurate end-to-end spotting of arbitrarily-shaped text. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.
- [173] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 10126–10135, 2020.
- [174] Yi Wang, Zhen-Peng Bian, Yunhao Zhou, and Lap-Pui Chau. Rethinking and designing a high-performing automatic license plate recognition approach. *IEEE Trans. Intell. Transp. Syst.*, 2021.
- [175] Changhao Wu, Shugong Xu, Guocong Song, and Shunqing Zhang. How many labeled license plates are needed? In *Chinese Conf. Patt. Recogn. Comp. Vis.*, pages 334–346. Springer, 2018.

- [176] Liang Wu, Chengquan Zhang, Jiaming Liu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. Editing text in the wild. In *Proc. ACM Int. Conf. Multimedia*, pages 1500–1508, 2019.
- [177] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4622–4630, 2016.
- [178] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets. *Comput. Vis. Image Underst.*, 163:21–40, 2017.
- [179] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019.
- [180] Linjie Xing, Zhi Tian, Weilin Huang, and Matthew R Scott. Convolutional character networks. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 9126–9136, 2019.
- [181] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1316–1324, 2018.
- [182] Zhenbo Xu, Wei Yang, Ajin Meng, Nanxue Lu, Huan Huang, Changchun Ying, and Liusheng Huang. Towards end-to-end license plate detection and recognition: A large dataset and baseline. In *Proc. Eur. Conf. Comp. Vis.*, pages 255–271, 2018.
- [183] Mingkun Yang, Yushuo Guan, Minghui Liao, Xin He, Kaigui Bian, Song Bai, Cong Yao, and Xiang Bai. Symmetry-constrained rectification network for scene text recognition. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 9147–9156, 2019.
- [184] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1083–1090, 2012.
- [185] Qixiang Ye and David Doermann. Text detection and recognition in imagery: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(7):1480–1500, 2014.

- [186] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 1031–1042, 2018.
- [187] Fei Yin, Yi-Chao Wu, Xu-Yao Zhang, and Cheng-Lin Liu. Scene text recognition with sliding convolutional character models. *arXiv preprint arXiv:1709.01727*, 2017.
- [188] Xu-Cheng Yin, Wei-Yi Pei, Jun Zhang, and Hong-Wei Hao. Multi-orientation scene text detection with adaptive clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(9):1930–1937, 2015.
- [189] Xu-Cheng Yin, Chun Yang, Wei-Yi Pei, Haixia Man, Jun Zhang, Erik Learned-Miller, and Hong Yu. DeTEXT: A database for evaluating text extraction from biomedical literature figures. *Plos one*, 10(5):e0126200, 2015.
- [190] SHI Yu, LI HaiYang, and Frank K Soong. A unified framework for symbol segmentation and recognition of handwritten mathematical expressions. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, volume 2, pages 854–858. IEEE, 2007.
- [191] Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. A large chinese text dataset in the wild. *JCST*, 34(3):509–521, 2019.
- [192] Fangneng Zhan, Shijian Lu, and Chuhui Xue. Verisimilar image synthesis for accurate detection and recognition of texts in scenes. In *Proc. Eur. Conf. Comp. Vis.*, pages 249–266, 2018.
- [193] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 5907–5915, 2017.
- [194] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1947–1962, 2018.

-
- [195] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23(10):1499–1503, 2016.
- [196] Linjiang Zhang, Peng Wang, Hui Li, Zhen Li, Chunhua Shen, and Yanning Zhang. A robust attentional framework for license plate recognition in the wild. *IEEE Trans. Intell. Transp. Syst.*, 2020.
- [197] Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. ICDAR2019 robust reading challenge on reading chinese text on signboard. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 1577–1581, 2019.
- [198] Sheng Zhang, Yuliang Liu, Lianwen Jin, and Canjie Luo. Feature enhancement network: A refined scene text detector. In *Proc. AAAI Conf. Artificial Intell.*, volume 32, 2018.
- [199] Shuaitao Zhang, Yuliang Liu, Lianwen Jin, Yaoxiong Huang, and Songxuan Lai. Ensnet: Ensconce text in the wild. In *Proc. AAAI Conf. Artificial Intell.*, volume 33, pages 801–808, 2019.
- [200] Xiaoci Zhang, Naijie Gu, Hong Ye, and Chuanwen Lin. Vehicle license plate detection and recognition using deep neural networks and generative adversarial networks. *J. Electronic imaging*, 27(4), 2018.
- [201] Ilia Zharikov, Philipp Nikitin, Ilia Vasiliev, and Vladimir Dokholyan. DDI-100: Dataset for text detection and recognition. In *ISCSIC*, pages 1–5, 2020.
- [202] Sergey Zherzdev and Alexey Gruzdev. Lprnet: License plate recognition via deep neural networks. *arXiv preprint arXiv:1806.10447*, 2018.
- [203] Zhuoyao Zhong, Lianwen Jin, and Shuangping Huang. DeepText: A new approach for text proposal generation and text detection in natural images. In *Proc. IEEE Int. Conf. Acoustics, Speech, & Signal Process.*, pages 1208–1212, 2017.
- [204] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5551–5560, 2017.
- [205] Yu Zhou, Hongtao Xie, Shancheng Fang, Jing Wang, Zhengjun Zha, and Yongdong Zhang. TDI TextSpotter: Taking data imbalance into account

- in scene text spotting. In *Proc. ACM Int. Conf. Multimedia*, pages 2510–2518, 2021.
- [206] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proc. ACM Int. Conf. Multimedia*, pages 4857–4866, 2022.
- [207] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 2223–2232, 2017.
- [208] Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, and Wayne Zhang. Fourier contour embedding for arbitrary-shaped text detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3123–3131, 2021.
- [209] Yongjie Zou, Yongjun Zhang, Jun Yan, Xiaoxu Jiang, Tengjie Huang, Haisheng Fan, and Zhongwei Cui. License plate detection and recognition based on yolov3 and ilprnet. *Signal, Image and Video Processing*, 16(2): 473–480, 2022.