IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 April 2024, accepted 15 April 2024, date of publication 30 April 2024, date of current version 9 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3395374

RESEARCH ARTICLE

SignboardText: Text Detection and Recognition in In-the-Wild Signboard Images

TIEN DO, THUYEN TRAN, THUA NGUYEN, DUY-DINH LE, (Member, IEEE), AND THANH DUC NGO

University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam

Corresponding author: Thanh Duc Ngo (thanhnd@uit.edu.vn)

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number DS2021-26-02.

ABSTRACT Scene text detection and recognition have attracted much attention in recent years because of their potential applications. Detecting and recognizing texts in images may suffer from scene complexity and text variations. Some of these problematic cases are included in popular benchmark datasets, but only to a limited extent. In this work, we investigate the problem of scene text detection and recognition in a domain with extreme challenges. We focus on in-the-wild signboard images in which text commonly appears in different fonts, sizes, artistic styles, or languages with cluttered backgrounds. We first contribute an in-the-wild signboard dataset with 79K text instances on both line-level and word-level across 2,104 scene images. We then comprehensively evaluated recent state-of-the-art (SOTA) approaches for text detection and recognition on the dataset. By doing this, we expect to realize the barriers of current state-of-the-art approaches to solving the extremely challenging issues of scene text detection and recognition, as well as their applicability in this domain. Code and dataset are available at https://github.com/aiclub-uit/SignboardText/ and IEEE DataPort.

INDEX TERMS Signboard images, scene text detection, scene text recognition.

I. INTRODUCTION

Image understanding is a topic of great interest in the research community, with numerous applications. Text in images is essential for naturally comprehending the images. Text can assist in extracting important information that is difficult to find by relying solely on scenery analysis. However, detecting and recognizing text in natural images remains a challenging problem due to scene complexity and text variations.

Popular publicly available datasets like COCO-Text [1], ICDAR 2015 [2], and Total-Text [3] are crucial in advancing scene text understanding research. They serve as platforms for difficulty discovery as well as benchmarks for evaluating advancement. Scene text detection and recognition have made significant strides lately, despite the datasets only partially reflecting the difficulties of the problem. In this work, we hope to advance understanding of text in unconstrained scenes by addressing a domain with extreme challenges. We focus on in-the-wild signboard images. Text on signboards may be displayed in a variety of fonts, sizes, artistic styles, or languages, depending on the signboard maker's or the artist's creativity to catch the attention of the viewer. Additionally, text instances on the same signboard may convey important semantic or design correlations. The ability to read text precisely in signboard images may assist in a variety of real-world applications, including geo-localization, autonomous driving, and disability assistance.

There are two main contributions of this work. First, we introduce an in-the-wild signboard dataset, named SignboardText, for scene text detection and recognition. The dataset consists of 79,814 manually annotated text instances at both line-level and word-level from 2,104 scene images. SignboardText includes a variety of words in different real-life conditions when the background is cluttered, the text is curved or distorted, the fonts are varied, or the art form is different (as shown in Figure 1). The signboard's typeface

The associate editor coordinating the review of this manuscript and approving it for publication was Michele Nappi.

FIGURE 1. Texts on signboards commonly appear in different fonts, sizes, artistic styles, or languages with cluttered backgrounds.

is the Latin alphabet, a widely used script worldwide, which combines with the aesthetic style of texts, creating a unique challenge for the dataset. We collect signboard images in different languages, with and without tone marks. The art-style texts combined with the system of tone marks, which is common in many language systems, are a prominent attribute of the dataset, leading to challenging detection and recognition cases of generic design methods.

Second, we provide a comprehensive evaluation and analysis of recent state-of-the-art (SOTA) methods for text detection and recognition on the SignboardText. We report the performance of different approaches, including TextSnake [4], DBNet [5], PSNet [6], PAN [7], DRRG [8], FCENet [9], and DPText-DETR [10];PaddleOCR [11], STAR-Net [12], SATRN [13], ViTSTR [14], ABINet [15], PARSeq [16], VietOCR,[1] Tesseract,[2] and also end-to-end methods such as ABCNet [17], ABCNet v2 [18], and DeepSolo [19]. The results and analyses of these methods serve as a valuable point of reference for researchers working on this specific problem.

The content of this paper is divided into the following sections: Section II examines a number of related works. Sections III and IV present the collected dataset as well as the analysis and evaluation based on the experiment results. Section V contains the conclusions.

II. RELATED WORKS

In this section, we first provide an overview of recent works on scene text detection and recognition. We categorize them mainly based on their approaches. Then, we summarize existing popular scene text datasets. Datasets are essential for training and evaluating proposed methods, enhancing understanding of the diversity and difficulty of scene text in images.

A. TEXT DETECTION AND RECOGNITION

Prior to the advent of deep learning, text detection and recognition methods relied heavily on low-level or mid-level handcrafted image features [20], [21], [22], [23], [24], [25], [26], which necessitated time-consuming and repeated pre-processing and post-processing procedures. Due to the limited representation capability of handcrafted features and the complexity of pipelines, such techniques are ill-equipped to deal with complicated situations.

[1]https://github.com/pbcquoc/vietocr

[2]https://github.com/tesseract-ocr/tesseract

TABLE 1. Number of signboard images from ICDAR2015, Total-Text, VinText, and manual collection.

Dataset	ICDAR2015	TotalText	VinText	Ours	Total
No. of sign board images	2	411	516	1,175	2,104
No. of text instances	20,261			59,588	79,849

Recent research on textual information extraction in images, as highlighted in [27], faces three primary challenges. The first includes text detection methods, which aim at localizing text instances in images. The second includes methods solely focusing on recognizing texts in given cropped text regions. The third takes both text localization and recognition into a single pipeline. Deep learning-based methods have emerged as the most advanced techniques in recent years.

1) TEXT DETECTION

Many algorithms are greatly influenced by and modelled after object detectors. In general, the development of scene text detection algorithms is divided into three stages [27]. Learning-based techniques are equipped with multi-step pipelines in the first phase to replace handcrafted features. Text center lines [28], [29] and single characters [30], [31] constitute key components of design. They are used to construct structures from the bottom up. In the second stage of development, generic object detection techniques are employed for the problem of scene text detection. Text detection methods for scenes are introduced by modifying detectors' region proposal and bounding box regression modules to locate text instances [32], [33], [34]. When dealing with irregular text, the performance of one-staged approaches is still bound by the receptive field's limitations. Two-staged approaches, on the other hand, are inefficient. Several works propose using sub-text components in the third stage to address long and irregular text problems. These methods use neural networks to predict local characteristics or segments and a post-processing step to reconstruct text instances at the pixel-level [35], [36], component-level [8], [34], [37], or character-level [38]. In the final step of post-processing, segments are combined into text instances. The use of sub-text components increases the flexibility and generality of text instance form and aspect ratio in detection.

2) TEXT RECOGNITION

Text recognition aims to translate a cropped text instance image into a target string sequence. There are two main categories of scene text recognition methods [39]: segmentation-based methods and segmentation-free methods.

Segmentation-based methods aim at predicting distinct pixel labels for each object instance. These methods can be roughly divided into top-down approaches and bottom-

up approaches. Top-down methods such as [40], [41], [42], and [43] aim to locate bounding boxes first, then in the second stage, segment the instances mask using within bounding boxes. [41] adopt VGG-16 as its backbone and combine and utilize low-level features (edge, color, and texture) and high-level features using multiple stages. Character bounding boxes are repeatedly predicted in multiple stages of the VGG-16, specifically. Then, the network outputs the final segmentation maps. Due to the assumption made by its prediction module (i.e., words are roughly sorted from left to right), [41] is not quite accurate in other scenarios. To address this disadvantage, [42] uses RNN for context modeling and a geometry branch to ensure characters are predicted in the correct order. Likewise, methods belong to MaskTextSpotter family [40], [43], integrates a spatial attention module, which helps complement the character segmentation sub-module by mitigating the limitation of lacking character-level annotations in the majority of the datasets. Segmentation-based methods [40], [41], [42], [43] usually include three steps: image preprocessing, character segmentation, and character recognition. Meanwhile, segmentation-free methods focus on directly mapping the whole text line or word into the target string using an encoder-decoder framework. Attention decoders are regarded as less adaptable than segmentation-based methods when it comes to identifying irregular text, including instances of oriented or curved text [42].

A typical segmentation-free method contains four main stages, including image preprocessing, feature representation, sequence modeling, and prediction. The purpose of image preprocessing, such as background removal, picture super-resolution, and rectification, is to enhance the quality of an image. By doing so, it has the potential to enhance the representation of features and improve recognition in subsequent stages. The utilization of Generative Adversarial Networks (GANs) for the purpose of enhancing the resolution of low-quality photos to a $2\times$ super-resolution image was demonstrated in the study conducted by Wang et al. [44]. They employed deformable attention and convolution, techniques also utilized in [45] and [46]. Other approaches, such as rectification, aim to normalize highly curved or distorted text instances. Several methods that fall under this category [12], [47], [48], [49], [50] use a variant of the spatial transformer network (STN) [51] to estimate spatial transformation via control points to correct the input image into a more straight shape. CNN networks are then commonly utilized at the feature representation stage to extract robust representations for the prediction stage. For instance, the VGG [52] network has been employed in [47] and [53]. Similarly, the ResNet [54] has been utilized in [12], [42], [49], [50], and [55]. Several works, including [14], [16], have recently adopted vision transformer as the backbone.

Typically, the sequence modeling stage is used to link visual features with predictions. It improves recognition by using bidirectional long short-term memory (BiLSTM) [56] to capture long-range dependencies in the visual features created in the previous stage and sending contextual cues to the prediction stage [12], [47], [49], [50], [55]. The fourth and final stage is prediction, in which the target string is predicted. Connectionist temporal classification (CTC) [57] and the attention mechanism [58] are two popular approaches used in this stage. The utilization of CTC as the prediction stage was initially introduced by [59], and subsequently, several prediction methods based on CTC, including [12], [60], and [61], have demonstrated remarkable performance. Visual attention, introduced in [58], has been integrated in recent works, improving recognition by identifying more informative and discriminative image regions. Examples of this include [47], [53], [55], [62], [63], [64], [65], [66], where feature selection and decoding were carried out using the attention mechanism. Vanilla attention was used in [55], [62], and [64]. To address attention drift, location-specific information was incorporated in [66]. In [66], the decoder decodes individual characters with a dynamic ratio between context and positional clues.

3) END-TO-END SCENE TEXT RECOGNITION

The goal of end-to-end scene text recognition (also referred to as text spotting) is to address both text detection and recognition simultaneously, as opposed to treating each as a separate task. According to [67] and [68], current text spotting techniques can be broadly divided into two categories: two-stage scene text spotters and single-shot scene text spotting methods.

a: TWO-STAGE SCENE TEXT SPOTTERS

The early scene text spotters, such as the Textboxes methods [69], [70], comprise two individual parts, i.e., the text detector and the recognizer, in a unified pipeline. The detection part of [69] and [70] uses SSD [71], while the recognition part uses CRNN [59]. Text detection and recognition are considered independent optimization tasks within the pipeline in these methods. Disjoint optimization, however, may make it difficult to identify the optimal solution. There are end-to-end techniques, such as [72], that jointly learn text detector and recognizer utilizing a curriculum learning model in order to mitigate the suboptimal relationship between text detection and recognition. Methods such as the MaskTextSpotter methods ([40], [43]) have recently adopted a segmentation approach. In [40] and [43], a Region-of-Interest (RoI) module is used to feed candidate regions into its Fast-RCNN branch to generate semantic segmentation maps. Feng et al. [73] use a sliding window method, RoISlide, inspired by the works of Long et al. [4], to read the text along the centerline of the text instances. To convert arbitrary-shape texts into conventional ones, the ABCNet based methods [17], [18] uses BezierAlign with learnable parameters.

b: ONE-STAGE TEXT SPOTTERS

One-stage text spotter methods [19], [74], [75] attempt to integrate the detector and recognizer into a one-stage network to avoid the adverse effects of RoI cropping. PGNet [74]

predicts text using center-point sequences. DeepSolo [19], taking inspiration from [17] and [18], devises a much simpler Bezier center curve proposal scheme and a novel query formulation. Finally, a simple linear projection can classify characters using these query cues.

B. SCENE TEXT DATASETS

Extracting textual information from images has caught the interest of the research community. Annual competitions on this problem have revealed that there are still many challenges to be solved. Recent research frequently focuses on solving problems posed in the Robust Reading competition [76]. Several popular datasets were released as a result of this contest, including COCO-Text [1], MSRA-TD500 [77], ICDAR 2013 [78], ICDAR 2015 [2], Total-Text [3], CTW1500 [79], VinText [80], for the problem of scene text detection and recognition, and ICDAR2017 [1], E2E-MLT [81], for multilingual scene text detection and language identification.

Language is a property related to text in images. ICDAR 2013 [78] and ICDAR 2015 [2] are well-known and popular datasets in English. These datasets emphasize small, oriented text. COCO-Text [1], which has a large number of images (63K), is another widely used dataset with texts in English. Texts in images are annotated at word level. Texts in COCO-Text, either machine-printed or hand-written, appear in different contexts, e.g., printed on objects such as baseball bats or backpacks. Total-Text [3] is another English dataset with 1,525 images for scene text. This dataset is primarily concerned with irregular text, particularly curved text. As a result, texts were annotated in different text orientations, including horizontal, multi-oriented, and curved.

In [82], the authors proposed a methodology that covers detection, orientation prediction, and recognition of Urdu ligatures in outdoor images. Urdu text is a cursive script that is part of the non-Latin family of scripts that includes Arabic, Chinese, and Hindi. This work introduced a dataset comprising 4.2K and 51K synthetic images embedded with Urdu text, generated using CLE annotation, and 1,094 real-world images with more than 12K Urdu characters. Utilizing the dataset, various methods were evaluated for different tasks, including detection, orientation prediction, and ligature recognition.

There are datasets with Chinese language, like CTW-1500 and RCTW-17. In CTW1500, the authors collected 1,500 images from the Internet and image libraries. All images are annotated with 10,751 cropped text instances, including 3,530 curved text. Another recent dataset is VinText, which focuses on Vietnamese scene text. This collection of images with text depicts objects from everyday life in Vietnam, such as clothing, books, storefronts, and street walls. VinText is the only and largest Vietnamese dataset at the moment, with 2,000 images and roughly 56,000 text instances. In general, most of the datsets are prepared for general scene text problem rather than specific domain like texts in signboard images. To the best of our knowledge, there is only one dataset related to signboard images introduced by Zhang et al [83]. The dataset is named ShopSign. It contains 25,770 images captured by smartphones. Images in ShopSign are manually annotated in a text-line-based manner. However, ShopSign solely contains Chinese texts, with the goal of furthering research in Chinese OCR.

Our dataset (namely SignboardText) contains Latin-based texts in two different languages, i.e., English and Vietnamese. Texts in both languages can appear simultaneously on the same signboard. We chose English to represent Latin-alphabet-based languages that do not have tone marks, while Vietnamese represents Latin-alphabet-based languages that do have tone marks. Tone marks are common in many languages and can lead to failure detection and recognition. Our dataset, when combined with ShopSign, provides a comprehensive benchmark for scene text detection and recognition in this new domain.

Selected samples of some of the aforementioned datasets are depicted in Figure 2, which can aid in the visualization of their characteristics. Table 2 is a statistics table that summarizes the datasets.

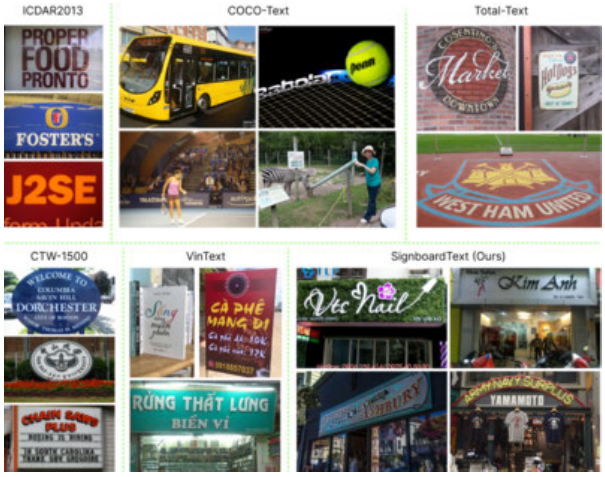

FIGURE 2. Sample images of some publicly available scene text datasets.

III. SignboardText

In this section, we provide a detailed description of our dataset, named SignboardText. It is made up of two parts: (1) 1,175 images manually labeled with a total of 59,588 text instances at the line and word levels (see Table 2); and (2) 929 signboard images collected from the VinText, Total-Text, and ICDAR15 datasets (as shown in Table 1). Each text instance in the first part of our dataset has a quadrilateral bounding box and a ground truth character sequence associated with it. In the second part, images are selected if they contain signboards (as illustrated in Figure 1). This portion of the dataset comprises 20,261 text instances at word levels. This bring the total text instances of our final dataset up into **79,814**. The dataset enhances the diversity and complexity of the text understanding in scene images, making it more representative of real-world scenarios. This

TABLE 2. Our **SignboardText** dataset and other public datasets for scene text detection and recognition. Our dataset can be compared with CTW-1500, Total-Text, and VinText in terms of quantity and diversity of text orientation. EN stands for english, CN for Chinese, VI for Vietnamese, UR for Urdu, and ML for multi-language. The term 'Regular' text refers to datasets in which the majority of text instances are simple, such as frontal and/or horizontal. The 'Irregular' collection contains the majority of text instances that are low-resolution, perspective warped, or curved.

	Datasets	Language	No. of Scene Images	Text Shape			Annotation Level		
				Horizontal	Arbitrary Quadrilateral	Multi-oriented	Character	Word	Line
Regular Text	RCTW-17/CTW-12K [84] *	CH	12,263	✓	✓	-	-	-	✓
	IC13 [78]	EN	462	✓	-	-	5,394	2,550	-
	CTW-1500 [79]	ML	1,500	✓	✓	✓	-	-	10,751
Irregular Text	COCO-Text [1]	EN	63,686	✓	✓	-	-	145,859	-
	CUTE80 [85]	EN	80	-	-	✓	-	-	466
	IC15 [2]	EN	1,500	✓	✓	-	-	17,548	-
	Total-Text [3]	EN	9,330	✓	✓	✓	-	1,525	-
	VinText [80]	VI	2,000	✓	✓	✓	-	56,000	-
	ShopSign [83]	CH	25,770	✓	✓	✓	-	-	196,010
	Urdu [82]	UR	1,094	✓	-	✓	12,000	-	-
	SignboardText (Ours)	ML	2,104	✓	✓	✓	-	68,899	10,950

dataset allows for a more robust evaluation of scene text recognition algorithms and their performance in handling new challenging issues, e.g., artistic-style text on signboards.

A. DATA CRAWLING AND ANNOTATION

In the initial phase of our data preparation, we collect images from Google Images using specific keywords that are related to billboards, traffic signs, hospital or school entrances, conferences, ceremonies, and other commonly encountered terms. These collected images were then categorized into different contexts, including "shop", "hospital", "school", "bank", "salon", and "other". Furthermore, we also classified the SignboardText dataset based on the specific locations where the images were captured, such as "fashion shop," "grocery store," "bakery," "food store," "coffee shop," and more. The distribution of our dataset's contexts can be observed in Figure 3. The diverse range of locations is one of the main factors contributing to the wide variety of fonts, styles, and backgrounds present in our dataset. Some representative samples from our dataset are illustrated in Figure 2.

To annotate the images, we divided them across 15 annotators and utilized the PPOCRLabel[3] tool for annotation. Following the ICDAR15 standard [2], we annotated each image with all of the text instances, polygons, and content

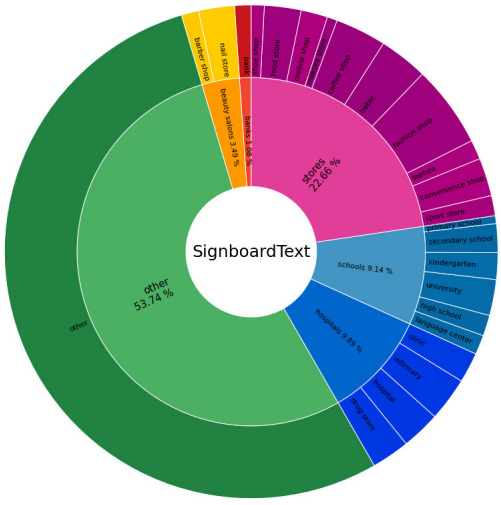

FIGURE 3. Categories of signboards in SignboardText.

that were present. Manual annotations were done on each and every image. After completing their workloads, each annotator uploaded their work to a shared storage location for cross-checking purposes. Team members view and cross-check each other's work, and their annotations are reviewed to ensure that all requirements have been met. This cross-checking process ensures the accuracy and consistency of the annotations. Before completing the annotations, the team

[3]https://github.com/Evezerest/PPOCRLabel

members discuss and resolve any discrepancies or errors that they find. Images with serious conflicting annotations are removed from the dataset, e.g., unreadable or hard-to-read text in poor imaging conditions.

B. DATASET STATISTICS

SignboardText contains 79,814 text instances with both line-level and word-level annotations for text detection and recognition. The annotation includes a wide variety of text types, including words, characters, digits, and also special characters such as dashes, dots, etc., as summarized in Table 5. In addition, taking into account an accented language, such as Vietnamese, the dataset includes new challenging issues with text detection and recognition regarding tone marks. The different tone marks in the same word can lead to different meanings, such as a, á, à, ã, ả, ạ. As a result, even missing or confusing a single tone mark in a word might result in fatal recognition results. The diversity of words with tone marks as well as background/text color, font styles, and font size in our dataset, SignboardText, can be seen in Figure 4, which serves as an example. The small size of the tone mark compared to the size of the whole word can be difficult for the recent scene text detection and recognition methods that mostly train on unaccented languages like English or Chinese. Besides, detecting text with tone marks may suffer because of the background. The tone mark makes the regions covered by a word much larger, hence including more background regions. This can result in decreased accuracy and increased false-positive rates for text detection algorithms, as they may struggle to differentiate between the actual text and the surrounding background. Moreover, the presence of tone marks can also introduce challenges in text recognition as it requires specialized algorithms that are capable of accurately interpreting and understanding the meaning behind these marks. Addressing these challenges is crucial for improving the performance of text detection and recognition algorithms in various applications.

FIGURE 4. Examples of texts cropped from signboard images.

As shown in Table 3, the majority of images in SignboardText have dimensions that are less than 1,000 pixels in both width and height. There are a few images with dimensions greater than 3,000 pixels. The presence of images with dimensions greater than 3,000 pixels indicates the potential variability in image sizes within the dataset. There are various word sizes in SignboardText. The mean width is 128 pixels, while the mean height is 60. However, the maximum width or height can be 20 times larger than the minimum ones. A similar observation is found with lines. This diversity in sizes of images, words, and lines poses great challenges for text detection and recognition algorithms, as they need to be able to handle images of different scales effectively.

TABLE 3. A summary of the sizes of images, annotated word boxes, and annotated line boxes.

	Image Size		Word Size		Line Size	
	Width	Height	Width	Height	Width	Height
Min	190	86	16	7	21	8
Max	4608	4160	3556	1125	3952	1032
Mean	1015.99	710.75	128.03	60.75	255.33	52.68

TABLE 4. A summary of the word length statistics of the annotated words in the dataset. The word lengths are divided into five different length categories.

Word Length	Percentage (%)
1 - 5	79.42
6 - 10	16.12
11 - 15	3.49
16 - 20	0.47
above 20	0.5

The average number of words per image is 40 (as shown in Figure 5). And words with 1–5 characters (word length) make up 79.42% of the annotated words. Most of the words (95.54%) have less than 10 characters. Long words are mainly related to email addresses or website addresses on signboards. A summary of word length statistics is shown in Table 4.Beside word length, Table 5 provides more detailed information on the types of characters that appear in words. Due to the nature of this domain, words do not solely contain alphabets; a large portion of words also include digits and special characters such as email addresses, social network links, websites, and telephone numbers. We also annotate signboard boxes with text that is unreadable in this dataset. The label '###' (unreadable) is used to indicate text in a box that is not readable by humans. Low resolution, small size, or deformed text are the main causes of unreadable texts. Texts that are unreadable can provide valuable information for scene analysis and understanding. They serve as crucial context clues to enhance other algorithms like image classification, object detection, or semantic segmentation. Therefore, even though they may be illegible to human readers, their presence in a scene is important for comprehensive analysis.

IV. EVALUATION AND ANALYSIS

We conducted an extensive evaluation of recent state-of-the-art (SOTA) methods on SignboardText in order to identify the limitations of current state-of-the-art approaches to solving the extremely challenging issues of scene text detection and recognition in this domain.

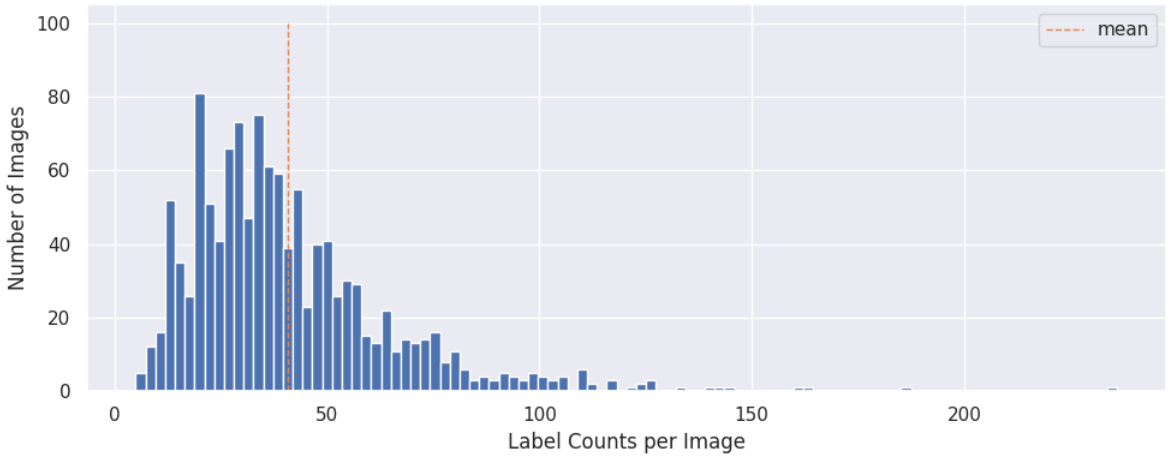

FIGURE 5. Number of words (i.e., labels) per image in SignboardText. The average number of words per image is 40.

TABLE 5. A summary of text instance types. When the text contained within a text box is not legible to human readers, they are designated with the notation '###' (unreadable). Most unreadable texts are due to low resolution, small size, or distorted text. In a category denoted as 'special characters', text comprise characters other than alphabets and digits such as '@' (found in emails), '-' and '.' (found in telephone numbers), '/' (found in Facebook addresses), and so forth, as shown on signboards.

Text Instances	Alphabet Only	Digits Only	Special Characters	Alphabet and Digits	Alphabet and Special Characters	Digits and Special Characters	Alphabet, Digits and Special Characters	Readable %
Lines	2,689	274	33	48	5,278	304	2,282	99.96
Words	46,691	3,413	1,584	647	4,651	2,066	553	86.63

A. EVALUATION METRICS

We select standard and popular metrics for evaluation. To evaluate text detection methods, we employ Precision, Recall, and H-Mean following the evaluation protocol called TedEval (Text Detection Evaluation) [86]. With regard to TedEval, detectors are evaluated via an instance-matching policy and a character-level scoring policy. TedEval is considered more suitable than other evaluation protocols for text detection and less sensitive to ground-truth quality.

We employ accuracy and Levenshtein distance to evaluate text recognition methods. Accuracy reflects the ratio of the number of correctly recognized words (or lines) to the ground truth. A prediction is considered correct if all characters of the word (or the line) match the character labels. Meanwhile, Levenshtein distance is a lexical similarity measure that identifies the distance between a pair of predictions and the ground truth. It basically relies on counting the number of single-character edits (i.e., insertions, deletions, or substitutions) required to change one word into another. This measure provides a more comprehensive evaluation of the system's performance beyond simple word matching.

We also analyze the speed and processing time of the models to assess their practical applicability. We use the frame per second (FPS) statistic to compare the speed of different models. Furthermore, we assess the resource requirements of deep learning-based methods, with a particular emphasis on the consumption of computational resources such as RAM and GPU usage.

TABLE 6. Detection results at the world-level of the SOTA methods with pretrained models on SignboardText. † denotes the best performance. ‡ indicates the second.

Method	Year	Recall	Precision	H-mean
TextSnake [4]	2018	57.06	60.67	58.56
PANet [7]	2019	71.92‡	78.98	75.14‡
PSENet [6]	2019	78.09†	82.06	79.94†
ABCNet v1 [17]	2020	64.48	76.26	69.40
DBNet [5]	2020	60.70	73.39	66.09
DRRG [8]	2020	50.82	71.60	59.22
FCENet [9]	2021	70.38	75.93	72.90
ABCNet v2 [18]	2021	64.32	77.66	69.95
DPText-DETR [10]	2023	60.46	90.56‡	72.07
DeepSolo [19]	2023	63.11	91.05†	74.20

B. DETECTION EVALUATION

We evaluate several recent state-of-the-art (SOTA) scene text detection methods on the SignboardText dataset. These methods include TextSnake [4], PANet [7], PSENet [6], Differentiable Binarization (DBNet) [5], DRRG [8], FCENet [9], and DPText-DETR [10], as well as the end-to-end methods from the ABCNet series, such as ABCNet v1 [17], ABCNet v2 [18], and DeepSolo [19]. The SignboardText dataset covers various challenges encountered in real-world signboard text detection. By evaluating these state-of-the-art

TABLE 7. Detection results at the world-level of the SOTA methods on various datasets, including **Ours** (SignboardText), ICDAR2015, and Total-Text. * denotes a reproduced result.

Method	SignboardText			ICDAR2015			Total-Text		
	Recall	Precision	H-mean	Recall	Precision	H-mean	Recall	Precision	H-mean
TextSnake [4]	57.06	60.67	58.56	84.90 [4]	80.40 [4]	82.60 [4]	74.50 [4]	82.70 [4]	78.40 [4]
PANet [7]	71.92	78.98	75.14	77.80 [7]	82.90 [7]	80.30 [7]	81.00 [7]	89.30 [7]	85.00 [7]
PSENet [6]	78.09	82.06	79.94	85.51 [6]	88.71 [6]	87.08 [6]	77.96 [6]	84.02 [6]	80.87 [6]
ABCNet v1 [17]	64.48	76.26	69.40	86.33*	88.76*	87.53*	50.48*	66.53*	57.40*
DBNet [5]	83.20	91.80	87.30	82.70 [5]	88.20 [5]	85.40 [5]	82.50 [5]	87.10 [5]	84.70 [5]
DRRG [8]	50.82	71.60	59.22	84.70 [8]	84.69 [8]	88.53 [8]	86.56 [8]	86.54 [8]	85.73 [8]
FCENet [9]	70.38	75.93	72.90	82.60 [9]	90.10 [9]	86.20 [9]	88.34 [9]	82.43 [9]	85.28 [9]
ABCNet v2 [18]	64.32	77.66	69.95	90.40 [18]	86.00 [18]	88.10 [18]	90.20 [18]	84.10 [18]	87.00 [18]
DPText-DETR [10]	60.46	90.56	72.07	90.93*	41.61*	57.09*	86.40 [10]	91.80 [10]	89.00 [10]
DeepSolo [19]	63.11	91.05	74.20	92.54 [19]	87.19 [19]	89.79 [19]	93.19 [19]	84.64 [19]	88.72 [19]

TABLE 8. Detection results at line-level of the SOTA methods with pretrained models on SignboardText. † denotes the best performance. ‡ indicates the second.

Method	Recall	Precision	H-mean
TextSnake [4]	86.66†	55.65	67.78
PANet [7]	76.89	80.43†	78.62†
PSENet [6]	82.77	21.20	33.75
ABCNet v1 [17]	69.00	78.47‡	73.43
DBNet [5]	71.87	20.73	32.18
DRRG [8]	38.91	57.84	46.52
FCENet [9]	82.81‡	70.81	76.25‡
ABCNet v2 [18]	67.59	78.45	72.62
DPText-DETR [10]	61.33	25.84	36.36
DeepSolo [19]	63.50	24.98	35.85

TABLE 9. Speed and GPU usage of the detection methods.† denotes the best performance. ‡ indicates the second.

Method	FPS	GPU (MB)
TextSnake [4]	1.66	2,447
PANet [7]	34.18†	2,065
PSENet [6]	7.14	825
ABCNet v1 [17]	3.41	2,142
DBNet [5]	13.41‡	847
DRRG [8]	1.23	1,844
FCENet [9]	8.06	2,310
ABCNet v2 [18]	2.51	1,223
DPText-DETR [10]	6.26	2,692
DeepSolo [19]	3.20	9,266

methods on this dataset, we aim to provide insights into their strengths and weaknesses, contributing to the advancement of text detection techniques for signboard images.

1) TextSnake [4]

In TextSnake, a text instance is described as a sequence of ordered, overlapping disks centered at symmetric axes, each of which is associated with a potentially variable radius and orientation. Such geometry attributes are estimated via a Fully Convolutional Network (FCN) model.

2) PANet [7]

Wang et al. have designed a lightweight segmentation framework consisting of two modules, namely the feature pyramid enhancement module (FPEM) and the feature fusion module (FFM), where FPEM generates scale-wise feature maps from an input image and FFM aggregates those multi-scale feature maps to generate the final feature map. Then, a pixel aggregation method is applied to predict text instances on a final feature map. The pixels of text instances are aggregated with the appropriate text kernels nearest to the corresponding text instances. This method yields high accuracy and efficiency due to its low-cost segmentation process.

3) PSENet [6]

This method specifies each text instance with multiple predicted segmentation areas, denoted "kernel" for simplification. Each kernel has the same shape as the original text representation but in different proportions. To get the final findings, the Breadth-First-Search (BFS)-based progressive scaling algorithm is used.

4) ABCNet [17]

In order to accurately localize oriented and curved text, Liu et al. introduce a new concise parametric representation of curved scene text using Bezier curves. They also proposed a new sampling method called Bezier Align, an advanced sampling method that gives better results than RoI sampling, which was proposed back in the early stages of Mask RCNN.

5) DBNet [5]

Liu et al. proposed integrating a module named Differentiable Binarization in the network to help the binary encoding in the image segmentation step faster; it also makes the detection process faster and more accurate.

6) DRRG [8]

The authors present an innovative local graph that bridges a text proposal model via Convolutional Neural Network (CNN) and a deep relational reasoning network via Graph Convolutional Network (GCN), making the network end-to-end trainable. Every text instance is divided into a series of small rectangular components, and the geometry attributes (e.g., height, width, and orientation) of the small components are estimated by a text proposal model.

7) FCENet [9]

In the method proposed by Yiqin et al., features extracted by the backbone (ResNet50 with DCN) and FPN are fed into the shared header to detect texts. In the header, the classification branch predicts both the heat maps of text regions and those of text center regions, which are pixel-wise multiplied, resulting in the classification score map. The regression branch predicts the Fourier signature vectors, which are used to reconstruct text contours via the Inverse Fourier transformation (IFT). Given the reconstructed text contours with corresponding classification scores, the final detected texts are obtained with non-maximum suppression (NMS).

8) DPText-DETR [10]

DPText-DETR, a Dynamic Point Text DEtection TRansformer network, as a solution to the limitations of existing Transformer-based methods for scene text detection. DPText-DETR addresses the issues of sub-optimal training efficiency and performance caused by coarse positional query modeling. It leverages explicit point coordinates to generate position queries and dynamically updates them progressively. The Enhanced Factorized Self-Attention module improves the spatial inductive bias of non-local self-attention by providing circular shape guidance to point queries. Additionally, a new positional label form is designed to improve detection robustness. DPText-DETR offers a concise and effective approach for scene text detection using dynamic point-based modeling.

9) DeepSolo [19]

DeepSolo is an innovative approach to text spotting that integrates text detection and recognition in a unified framework, taking inspiration from the methodology employed in DETR [87]. It achieves this by employing a single Decoder with Explicit Points Solo, enabling simultaneous processing of both tasks. Text instances are represented as ordered points and modeled using learnable explicit point queries. These queries, which are sampled using Bezier curves as introduced in [17] and [18], are then decoded to extract crucial information such as the center line, boundary, script, and confidence of the text. This novel technique demonstrates promising potential in the field of text spotting.

Table 6 presents detection results at word-levels of the evaluated methods with their pretrained models on SignboardText. The main goal of these experiments is to realize the performance of these SOTA methods in this new domain without fine-tuning. By using precision, recall, and H-mean simultaneously, we may broadly observe their strengths and weaknesses. In terms of precision, we learn that the most recent works, i.e., DPTExt-DETR and DeepSolo, have the best performance, as they take the top two precision scores with over 90%. Without retraining in a new domain, their performance is remarkable. This observation is also reasonable since they are carefully designed to deal with irregular texts at the word-level and irregular texts take up a large portion of the text on signboards. However, their trade-off on recall is significant, as they missed approximately 40% of the words appearing on the signboards. In some applications, this level of recall may lead to a failure in content extraction and understanding. In terms of recall, PSENet and PANet achieve the best performance with 78.09% and 71.92%, respectively. Besides, they also achieve equivalent precision with 82.06% and 78.98%. In addition to their remarkable performance in terms of precision and recall, PSANet and PANet also excel in terms of H-mean. This metric takes into account both precision and recall to provide an overall measure of a model's effectiveness. With their well-balanced levels between precision and recall, PSANet and PANet emerge as the top two methods in terms of H-mean.

Table 7 provides a comprehensive evaluation of the SOTA detection methods at word-level across different datasets. The results reveal that the methods' overall performance on SignboardText, specifically the H-mean, is significantly lower than their performances on ICDAR2015 and Total-Text. This discrepancy can be attributed to the fact that diacritics, tonal marks, and artistic design from non-English texts made a significant impact on the performance of pretrained models that were originally trained in English settings.

In Table 9, we present the speed and GPU memory usage of the evaluated methods. Due to its lightweight segmentation framework, PANet is recognized as the fastest, with 34.18 frames per second. PSENet, with the advantage of the segmentation-based method, proves more powerful at word-level detection (84.29% H-mean). This method can be efficient for splitting close instances of text and detecting cases of text with arbitrary shapes. By relying on a Breadth-First-Search (BFS) based progressive scaling algorithm to formulate the final results, this results in a low-speed inference process (7.14 FPS). PANet has both the advantages of high speed (34.18 FPS) and average accuracy (78.62% H-mean at line level) since it is equipped with a low computational-cost segmentation head and learnable post-processing. However, it requires a large amount of resources (2065 MB of GPU memory). This can be problematic since many embedding systems are resource-constrained.

Table 8 summarizes the detection performance at line-levels of the methods on SingboardText. Compared to word-level detection, line-level detection is more sensitive to background noise and aspect ratios of detection boxes. PANet achieves the highest precision on line-level detection with 80.43%. And ABCNet v1 is followed up with 78.45%. The two most recent methods, DPTExt-DETR and DeepSolo, are among the methods with the lowest precision. This is because DPTExt-DETR and DeepSolo are not intentionally designed to deal with line detection. In terms of recall, TextSnake and FCENet are the two best methods. Overall, PANet and FCENet demonstrate the strongest performance in terms of both precision and recall, as indicated by their high H-mean

scores of 78.62% and 76.52%, respectively. This suggests that these methods are effective in accurately detecting lines while also minimizing false positives. However, it is worth noting that PANet achieves significantly higher precision compared to FCENet, indicating its superior ability to accurately identify line-level detections.

C. RECOGNITION EVALUATION

For recognition evaluation, we use a training set of 7,850,000 images generated by SynthText and 530,000 images generated using TextRecognitionDataGenerator. By this, we expect that the models can learn with large variations of text to be able to deal with challenging cases in SignboardText. We chose evaluate recent SOTA methods which have shown remarkable performance on other standard benchmarks.

1) DeepText [88]

The framework consists of four phases aimed at text recognition. The first step involves transforming text data of various shapes into straight text to facilitate processing. This transformed information is then passed through a convolutional neural network to extract relevant features (known as the feature extraction phase). Subsequently, the extracted features are fed into a model that learns and employs the relationship between sequences of characters (known as the sequence modeling phase). Finally, the framework predicts the characters present in the image to be recognized (known as the prediction phase).

2) PaddleOCR [11]

Researchers from Baidu have proposed the PaddleOCR architecture, which is based on upgraded, lightweight neural networks. The proposed framework consists of three main modules: text border detection, text corner correction, and text recognition. All three modules employ lightweight backbone networks to improve computational efficiency and make the method suitable for embedded applications. The first module employs a segmentation network-based text detector whose objective is to identify and segment the area of the image that contains the text. To correct the detection box, a geometric transformation is given to the image area in the second module. A convolutional recurrent neural network (CRNN) is utilized in the last stage, text recognition, to identify the text in the rectified bounding box.

3) STAR-Net [12]

STAR-Net introduces a Spatial Transformer Network (STN) to rectify text areas, making them more suitable for recognition. This approach is particularly effective at handling curved or distorted text, proving to be a valuable tool in scenarios where scene text is not perfectly aligned or formatted.

VietOCR[4] is a method proposed for recognizing Vietnamese handwritten and optical characters. It includes two main models: AttentionOCR, which is a combination of CNN architecture and attention seq2seq architecture, and TransformerOCR, which is a combination of CNN architecture and Transformer architecture.

[4]https://github.com/pbcquoc/vietocr

4) SAR [90]

SAR presents a simple yet efficient method for recognizing irregular text. SAR combines a sequence-to-sequence model with an attention mechanism, where 2D vector maps are used as input to an attention module, resulted in ''glimpse'' vectors (as one of the input of LSTM module).

5) SATRN [13]

SATRN makes use of the self-attention mechanism to investigate the 2D spatial relationships among characters. In combination with innovative 2D positional encoding, the encoder of SATRN mitigates the absence of crucial location information resulting from self-attention. As a result, this approach is capable of managing extreme cases, such as severe distortion.

FIGURE 6. Examples of signboards with artistic text. Texts with artistic styles are extremely common on signboards.

6) ViTSTR [14]

ViTSTR is a simple yet efficient implementation of vision transformers. It comprises only 12 identical encoder blocks without a decoder, and its prediction layer is a basic linear layer for projecting encoded features into predictions. Additionally, the authors improve the accuracy of ViTSTR through the application of diverse and multiple data augmentation techniques.

7) ABINet [15]

ABINet employs a language model to uncover the connection between visual and textual information. Furthermore, it introduces a bidirectional cloze network (BCN) to create feature representations that consider information from both the left and right directions. In addition, ABINet utilizes iterative refinements, starting with predictions from the vision model and then refining them through iterations using the language model.

TABLE 10. Recognition results of the evaluated SOTA methods on SignboardText. MJ, ST denotes MJSynth, SynthText respectively. † denotes the best performance. ‡ indicates the second.

Method	Year	Training data	Accuracy (%) (case-insensitive)
CRNN [59]	2015	MJ+ST	39.04
STAR-Net [12]	2016	MJ+ST	47.40
Rosetta [89]	2018	MJ+ST	46.12
SAR [90]	2019	MJ+ST	57.96
VietOCR-TranformerOCR	2019	MJ+ST	68.57
VietOCR-TranformerOCR	2019	VietOCR privated data	76.84 †
VietOCR-AttentionOCR	2019	MJ+ST	68.42
VietOCR-AttentionOCR	2019	VietOCR privated data	74.94
SATRN [13]	2020	MJ+ST	56.34
Paddle OCR	2020	MJ+ST	11.56
Paddle OCR	2020	VietOCR provided data	22.42
ViTSTR [14]	2021	MJ+ST	67.13
ABINet [15]	2021	MJ+ST	66.07
PARSeq [16]	2022	COCO-Text, RCTW17,Uber-Text, ArT, LSVT, MLT19, and ReCTS	68.55 ‡
SVTR [91]	2022	MJ+ST	52.23

TABLE 11. Recognition results of the evaluated methods on various datasets, including **Ours** (SignboardText), ICDAR2013, ICDAR2015, Total-Text, and COCO-Text. The ICDAR2013, ICDAR2015, Total-Text, and COCO-Text consist of 1,095, 2,077, 2,201, and 9,837 images, respectively, as outlined in the conventions introduced by DeepText [88]. * denotes a reproduced result. A citation on the right of a result indicates that the result is derived from the cited papers.

Method	SignboardText	ICDAR2013	ICDAR2015	ToTal-Text	COCO-Text
CRNN [59]	39.04	89.20 [88]	64.20 [88]	49.48*	32.46*
STAR-Net [12]	47.40	91.50 [88]	70.30 [88]	35.07*	24.14*
Rosetta [89]	46.12	89.00 [88]	66.00 [88]	15.81*	9.30*
SAR [90]	57.96	94.00 [90]	78.8 [90]	56.88*	66.80 [90]
Paddle OCR	11.56	94.09*	69.96*	66.20*	51.30*
VietOCR-TranformerOCR	68.57	89.85*	60.13*	56.07*	39.59*
SATRN [13]	56.34	94.10 [13]	79.00 [13]	57.38*	23.60*
ViTSTR [14]	67.13	94.20 [16]	78.70 [16]	58.47*	56.40 [16]
ABINet [15]	66.07	95.00 [16]	79.10 [16]	77.37*	57.10 [16]
PARSeq [16]	68.55	96.20 [16]	82.90 [16]	71.51*	64.00 [16]

Ground Truth TextSnake PANet

FIGURE 7. Examples to illustrate that the detection methods are confused by the artfully designed parts of the text.

8) PARSeq [16]

PARSeq leverages permutation language modeling to train an ensemble of internal autoregressive language models with shared weights. This approach combines context-aware AR (auto regressive) inference and context-free non-AR, along with iterative refinement using bidirectional context (iterative refinement is introduced in ABINet [15]). By doing so, PARSeq provides a unified solution that overcomes the shortcomings of traditional AR models and improves the overall accuracy and efficiency of context-aware STR.

9) SVTR [91]

SVTR represents another vision-only model that contextualizes dependencies at both the local level (such as stroke-like features) and the global level (such as inter-character features). This method employs a significantly simpler architecture compared to other methods that fall under the encoder-decoder category, like those mentioned in [13] and [15]. Consequently, it reduces both the computational time and resource requirements.

FIGURE 8. Examples of threshold sensitivity. With the IoU threshold equal to 0.5, the detected box does not cover the tone mark, thus leading to the wrong recognition. If the IoU is set to 0.7, the box is rejected. This may cause a drop in the recall.

TABLE 12. Speed and GPU usage comparison of recognition methods. * denotes the models that belong to the DeepText framework.

Method	FPS	GPU (MB)
CRNN [59]	79.55	283
STAR-Net [12]	57.03	1,527
Rosetta [89]	81.67	1,491
SAR [90]	5.33	1,530
Paddle OCR	532.25	297
VietOCR-TranformerOCR	5.57	1,433
VietOCR-AttentionOCR	28.80	1,319
SATRN [13]	2.30	1,502
ViTSTR [14]	10.74	1,238
ABINet [15]	4.63	1,356
PARSeq [16]	8.24	1,268

With the evaluation results shown in Table 10, VietOCR-TransformerOCR achieves the highest accuracy at 76.84%, while PARSeq comes in second with an accuracy of 68.55%, trailing VietOCR. The reason for this performance difference is that both VietOCR and PARSeq are trained on real-life datasets, which have distributions that closely match the target data. In contrast, synthetic datasets like MJSynth and SynthText have different characteristics. Table 12 summarizes the speed and GPU usage of the evaluated methods.

In Table 11, we summarize the recognition results of the evaluated methods across datasets, including SignboardText, ICDAR2013, ICDAR2015, Total-Text, and COCO-Text. Compared to ICDAR2013 and ICDAR2015, the performance of the methods significantly drops as they are applied to SignboardText. This indicates that SignboardText covers more challenging cases than ICDAR2013 and ICDAR2015, where the majority of texts are predominantly horizontal (as observed in Table 2). However, as we observed in the group of three datasets, including SignboardText, Total-Text, and COCO-Text, the performances of the evaluated methods are at the same low level. SignboardText demonstrates itself as a challenging dataset for SOTA text recognition methods. It is also worth noting that the size of SignboardText is smaller than that of the other two datasets. Hence, its source of challenge is not mainly scale related issues. This lower accuracy is attributed to the presence of unique challenges in the signboard domain.

We observe that one of the main challenges of detecting and recognizing texts from signboard images is related to texts with artistic styles (examples shown in Figure 6). Designers usually use artistic text on signboards to create a visual impact and attract attention. Artistic text can also convey a message or a mood that is related to the brand or the product. For example, some designers use typography that has a distinctive identity and style. Typography is the art of arranging typefaces in various combinations of font, size, and spacing. Some designers also use text art as a form of expression and creativity.

Beside the extreme variations of text instances with different art styles, tone marks in text are another challenging issue. Tone marks may cause the text recognizer to be confused with other artfully designed parts of the text, which are not helpful in recognizing the text. And bounding boxes of words with tone marks usually cover a larger background. This injects more noise from background regions into the learning models. Larger bounding boxes (to cover tone mark regions) also cause text detection results to be more sensitive to the IoU threshold. Tightly detected boxes are easily considered failure detection. Some examples are given in Fig. 7 and Fig. 8.

Although handling artistic text is challenging, it is important for scene understanding. Art text recognition on signboards can improve image understanding and analysis by providing additional data that can be used to train or test the models for these tasks. For example, art text recognition can help identify the location of a signboard in an image and use it as a reference point for other tasks. Art text recognition can also help to extract the text content from signboards and use it as a source of information for other tasks.

Moreover, the texts in signboard images are obtained from the real world, and as such, they are subject to direct influence from different factors such as low resolution, blurring, multiscale, diverse density, multilinguality, and invisibility. In these situations, text recognition is another issue that garners community interest.

V. CONCLUSION

We present SignboardText, an in-the-wild signboard dataset, for scene text detection and recognition. The dataset comprises 2,104 images, of which 79,814 instances of text

have been manually annotated at the word and line levels. SignboardText incorporates a diverse range of words to represent various real-life scenarios, such as those involving a cluttered background, curved or distorted text, varied typefaces, or distinct artistic styles. The art-style texts combined with the system of tone marks, which is common in many language systems, are a prominent attribute of the dataset, leading to challenging detection and recognition cases of generic design methods. We comprehensively evaluated recent state-of-the-art (SOTA) approaches for text detection and recognition on the dataset. Based on experimental results and analysis, we exposed the barriers of current state-of-the-art approaches to solving the extremely challenging issues of scene text detection and recognition in this new domain.

REFERENCES

[1] R. Gomez, B. Shi, L. Gomez, L. Numann, A. Veit, J. Matas, S. Belongie, and D. Karatzas, "ICDAR2017 robust reading challenge on COCO-text," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 01, Nov. 2017, pp. 1435–1443.

[2] X. Zhou, S. Zhou, C. Yao, Z. Cao, and Q. Yin, "ICDAR 2015 text reading in the wild competition," 2015, *arXiv:1506.03184*.

[3] C. K. Chng and C. S. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," 2017, *arXiv:1710.10400*.

[4] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "TextSnake: A flexible representation for detecting text of arbitrary shapes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 20–36.

[5] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time scene text detection with differentiable binarization," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11474–11481.

[6] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, and S. Shao, "Shape robust text detection with progressive scale expansion network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9328–9337.

[7] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu, G. Yu, and C. Shen, "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8439–8448.

[8] S.-X. Zhang, X. Zhu, J.-B. Hou, C. Liu, C. Yang, H. Wang, and X.-C. Yin, "Deep relational reasoning graph network for arbitrary shape text detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9696–9705.

[9] Y. Zhu, J. Chen, L. Liang, Z. Kuang, L. Jin, and W. Zhang, "Fourier contour embedding for arbitrary-shaped text detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3122–3130.

[10] M. Ye, J. Zhang, S. Zhao, J. Liu, B. Du, and D. Tao, "DPText-DETR: Towards better scene text detection with dynamic points in transformer," in *Proc. Conf. Artif. Intell. (AAAI)*, 2023, pp. 3241–3249.

[11] Y. Du, C. Li, R. Guo, X. Yin, W. Liu, J. Zhou, Y. Bai, Z. Yu, Y. Yang, Q. Dang, and H. Wang, "PP-OCR: A practical ultra lightweight OCR system," 2020, *arXiv:2009.09941*.

[12] W. Liu, C. Chen, K.-Y. Wong, Z. Su, and J. Han, "STAR-Net: A SpaTial attention residue network for scene text recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2016, p. 7.

[13] J. Lee, S. Park, J. Baek, S. J. Oh, S. Kim, and H. Lee, "On recognizing texts of arbitrary shapes with 2D self-attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2326–2335.

[14] R. Atienza, "Vision transformer for fast and efficient scene text recognition," in *Proc. Int. Conf. Document Anal. Recognit.* Cham, Switzerland: Springer, 2021, pp. 319–334.

[15] S. Fang, H. Xie, Y. Wang, Z. Mao, and Y. Zhang, "Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7094–7103.

[16] D. Bautista and R. Atienza, "Scene text recognition with permuted autoregressive sequence models," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 178–196, doi: 10.1007/978-3-031-19815-1_11.

[17] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, and L. Wang, "ABCNet: Real-time scene text spotting with adaptive Bezier-curve network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9806–9815.

[18] Y. Liu, C. Shen, L. Jin, T. He, P. Chen, C. Liu, and H. Chen, "ABCNet v2: Adaptive Bezier-curve network for real-time end-to-end text spotting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8048–8064, Nov. 2022.

[19] M. Ye, J. Zhang, S. Zhao, J. Liu, T. Liu, B. Du, and D. Tao, "DeepSolo: Let transformer decoder with explicit points solo for text spotting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19348–19357.

[20] J. Rodriguez and F. Perronnin, "Label embedding for text recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2013, pp. 1–5.

[21] C. Yao, X. Bai, B. Shi, and W. Liu, "Strokelets: A learned multi-scale representation for scene text recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 4042–4049.

[22] S. Lee and J. H. Kim, "Integrating multiple character proposals for robust scene text extraction," *Image Vis. Comput.*, vol. 31, no. 11, pp. 823–840, Nov. 2013.

[23] Q. Ye, W. Gao, W. Wang, and W. Zeng, "A robust text detection algorithm in images and video frames," in *Proc. 4th Int. Conf. Inf., Commun. Signal Process., 4th Pacific Rim Conf. Multimedia Joint*, 2003, pp. 802–806.

[24] P. Shivakumara, S. Bhowmick, B. Su, C. L. Tan, and U. Pal, "A new gradient based character segmentation method for video text recognition," in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 2011, pp. 126–130.

[25] K. Sheshadri and S. Divvala, "Exemplar driven character recognition in the wild," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–10.

[26] A. Mishra, K. Alahari, and C. Jawahar, "Scene text recognition using higher order language priors," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–25, doi: 10.5244/c.26.127.

[27] F. Naiemi, V. Ghods, and H. Khalesi, "Scene text detection and recognition: A survey," *Multimedia Tools Appl.*, vol. 81, no. 14, pp. 20255–20290, Jun. 2022.

[28] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4159–4167.

[29] D. He, X. Yang, C. Liang, Z. Zhou, A. G. Ororbia, D. Kifer, and C. L. Giles, "Multi-scale FCN with cascaded instance aware segmentation for arbitrary oriented word spotting in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 474–483.

[30] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. L. Tan, "Text flow: A unified text detection system in natural scene images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4651–4659.

[31] A. V. Goldberg, "An efficient implementation of a scaling minimum-cost flow algorithm," *J. Algorithms*, vol. 22, no. 1, pp. 1–29, Jan. 1997.

[32] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3066–3074.

[33] L. Yuliang, J. Lianwen, Z. Shuaitao, and Z. Sheng, "Detecting curve text in the wild: New dataset and new solution," 2017, *arXiv:1712.02170*.

[34] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3482–3490.

[35] D. Deng, H. Liu, X. Li, and D. Cai, "PixelLink: Detecting scene text via instance segmentation," 2018, *arXiv:1801.01315*.

[36] Y. Wu and P. Natarajan, "Self-organized text detection with minimal post-processing via border learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5010–5019.

[37] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 56–72.

[38] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9357–9366.

[39] X. Chen, L. Jin, Y. Zhu, C. Luo, and T. Wang, "Text recognition in the wild: A survey," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–35, 2021.

[40] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 67–83.

[41] M. Liao, J. Zhang, Z. Wan, F. Xie, J. Liang, P. Lyu, C. Yao, and X. Bai, "Scene text recognition from two-dimensional perspective," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8714–8721, doi: 10.1609/aaai.v33i01.33018714.

[42] Z. Wan, M. He, H. Chen, X. Bai, and C. Yao, "TextScanner: Reading characters in order for robust scene text recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12120–12127, doi: 10.1609/aaai.v34i07.6891.

[43] M. Liao, G. Pang, J. Huang, T. Hassner, and X. Bai, "Mask TextSpotter v3: Segmentation proposal network for robust scene text spotting," 2020, *arXiv:2007.09482*.

[44] W. Wang, E. Xie, P. Sun, W. Wang, L. Tian, C. Shen, and P. Luo, "TextSR: Content-aware text super-resolution guided by recognition," 2019, *arXiv:1909.07113*.

[45] L. Deng, Y. Gong, X. Lu, X. Yi, Z. Ma, and M. Xie, "Focus-enhanced scene text recognition with deformable convolutions," in *Proc. IEEE 5th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2019, pp. 1685–1689. [Online]. Available: https://api.semanticscholar.org/CorpusID:201670529

[46] W. Yang, J. Wu, J. Zhang, K. Gao, R. Du, Z. Wu, E. Firkat, and D. Li, "Deformable convolution and coordinate attention for fast cattle detection," *Comput. Electron. Agricult.*, vol. 211, Aug. 2023, Art. no. 108006. [Online]. Available: https://api.semanticscholar.org/CorpusID:259643672

[47] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4168–4176.

[48] S. Long, Y. Guan, B. Wang, K. Bian, and C. Yao, "Rethinking irregular scene text recognition," 2019, *arXiv:1908.11834*.

[49] F. Zhan and S. Lu, "ESIR: End-to-end scene text recognition via iterative image rectification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2054–2063.

[50] M. Yang, Y. Guan, M. Liao, X. He, K. Bian, S. Bai, C. Yao, and X. Bai, "Symmetry-constrained rectification network for scene text recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9146–9155.

[51] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," 2015, *arXiv:1506.02025*.

[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[53] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou, "Focusing attention: Towards accurate text recognition in natural images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5086–5094.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[55] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "ASTER: An attentional scene text recognizer with flexible rectification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2035–2048, Sep. 2019.

[56] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 855–868, May 2009.

[57] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[58] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[59] B. Shi, X. Bai, and C. 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.*, vol. 39, no. 11, pp. 2298–2304, Jun. 2016.

[60] Z. Wan, F. Xie, Y. Liu, X. Bai, and C. Yao, "2D-CTC for scene text recognition," 2019, *arXiv:1907.09705*.

[61] W. Hu, X. Cai, J. Hou, S. Yi, and Z. Lin, "GTC: Guided training of CTC towards efficient and accurate scene text recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11005–11012.

[62] C.-Y. Lee and S. Osindero, "Recursive recurrent nets with attention modeling for OCR in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2231–2239.

[63] X. Yang, D. He, Z. Zhou, D. Kifer, and C. L. Giles, "Learning to read irregular text with attention mechanisms," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, p. 3.

[64] C. Luo, L. Jin, and Z. Sun, "MORAN: A multi-object rectified attention network for scene text recognition," *Pattern Recognit.*, vol. 90, pp. 109–118, Jun. 2019.

[65] R. Litman, O. Anschel, S. Tsiper, R. Litman, S. Mazor, and R. Manmatha, "SCATTER: Selective context attentional scene text recognizer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11959–11969.

[66] X. Yue, Z. Kuang, C. Lin, H. Sun, and W. Zhang, "RobustScanner: Dynamically enhancing positional clues for robust text recognition," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 135–151.

[67] H. Li, P. Wang, and C. Shen, "Towards end-to-end text spotting with convolutional recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5248–5256.

[68] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "FOTS: Fast oriented text spotting with a unified network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5676–5685.

[69] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: A fast text detector with a single deep neural network," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2017, pp. 1–7, doi: 10.1609/aaai.v31i1.11196.

[70] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3676–3690, Aug. 2018.

[71] W. Liu, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf.* Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 21–37.

[72] M. Busta, L. Neumann, and J. Matas, "Deep TextSpotter: An end-to-end trainable scene text localization and recognition framework," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2231.

[73] W. Feng, W. He, F. Yin, X.-Y. Zhang, and C.-L. Liu, "TextDragon: An end-to-end framework for arbitrary shaped text spotting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9075–9084.

[74] P. Wang, C. Zhang, F. Qi, S. Liu, X. Zhang, P. Lyu, J. Han, J. Liu, E. Ding, and G. Shi, "PGNet: Real-time arbitrarily-shaped text spotting with point gathering network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, 2021, pp. 2782–2790, doi: 10.1609/aaai.v35i4.16383.

[75] X. Zhang, Y. Su, S. Tripathi, and Z. Tu, "Text spotting transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9509–9518.

[76] (2010). *ICDAR Robust Reading Competition (RRC)*. Accessed: Sep. 30, 2010. [Online]. Available: https://rrc.cvc.uab.es/

[77] J. Liu, C. Zhang, Y. Sun, J. Han, and E. Ding, "Detecting text in the wild with deep character embedding network," 2019, *arXiv:1901.00363*.

[78] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. Bigorda, S. Mestre, J. Mas, D. Mota, J. Almazan, and de las Heras, "Robust reading competition," in *Proc. Int. Conf. Document Anal. Recognit.*, 2013, pp. 1484–1493.

[79] Y. Liu, L. Jin, S. Zhang, C. Luo, and S. Zhang, "Curved scene text detection via transverse and longitudinal sequence connection," *Pattern Recognit.*, vol. 90, pp. 337–345, Jun. 2019.

[80] N. Nguyen, T. Nguyen, V. Tran, M.-T. Tran, T. D. Ngo, T. Huu Nguyen, and M. Hoai, "Dictionary-guided scene text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7379–7388.

[81] M. Bušta, Y. Patel, and J. Matas, "E2E-MLT—An unconstrained end-to-end method for multi-language scene text," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 127–143.

[82] S. Y. Arafat and M. J. Iqbal, "Urdu-text detection and recognition in natural scene images using deep learning," *IEEE Access*, vol. 8, pp. 96787–96803, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219313164

[83] C. Zhang, G. Peng, Y. Tao, F. Fu, W. Jiang, G. Almpanidis, and K. Chen, "ShopSign: A diverse scene text dataset of Chinese shop signs in street views," 2019, *arXiv:1903.10412*.

[84] B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. Belongie, S. Lu, and X. Bai, "ICDAR2017 competition on reading Chinese text in the wild (RCTW-17)," 2017, *arXiv:1708.09585*.

[85] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Syst. Appl.*, vol. 41, no. 18, pp. 8027–8048, Dec. 2014.

[86] C. Y. Lee, Y. Baek, and H. Lee, "TedEval: A fair evaluation metric for scene text detectors," in *Proc. Int. Conf. Document Anal. Recognit. Workshops (ICDARW)*, Sep. 2019, pp. 14–17.

[87] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[88] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, and H. Lee, "What is wrong with scene text recognition model comparisons? Dataset and model analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4714–4722. [Online]. Available: https://api.semanticscholar.org/CorpusID:102481180

[89] F. Borisyuk, A. Gordo, and V. Sivakumar, "Rosetta: Large scale system for text detection and recognition in images," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 71–79.

[90] H. Li, P. Wang, C. Shen, and G. Zhang, "Show, attend and read: A simple and strong baseline for irregular text recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8610–8617.

[91] Y. Du, Z. Chen, C. Jia, X. Yin, T. Zheng, C. Li, Y. Du, and Y.-G. Jiang, "SVTR: Scene text recognition with a single visual model," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 884–890, doi: 10.24963/ijcai.2022/124.

[92] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1480–1500, Jul. 2015.

[93] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 161–184, Jan. 2021.

[94] V. Lavrenko, T. M. Rath, and R. Manmatha, "Holistic word recognition for handwritten historical documents," in *Proc. 1st Int. Workshop Document Image Anal. Libraries*, Jan. 2004, pp. 278–287.

[95] M. Seuret, M. Alberti, M. Liwicki, and R. Ingold, "PCA-initialized deep neural networks applied to document image analysis," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, Nov. 2017, pp. 877–882.

[96] N. Arvanitopoulos and S. Süsstrunk, "Seam carving for text line extraction on color and grayscale historical manuscripts," in *Proc. 14th Int. Conf. Frontiers Handwriting Recognit.*, Sep. 2014, pp. 726–731.

[97] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*.

[98] Y. Liu and L. Jin, "Deep matching prior network: Toward tighter multi-oriented text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3454–3461.

[99] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.

[100] L. Rong, E. MengYi, L. JianQiang, and Z. HaiBin, "Weakly supervised text attention network for generating text proposals in scene images," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, Nov. 2017, pp. 324–330.

[101] M. Liao, Z. Zhu, B. Shi, G.-s. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5909–5918.

[102] S. Zhang, Y. Liu, L. Jin, and C. Luo, "Feature enhancement network: A refined scene text detector," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, pp. 2612–2619, doi: 10.1609/aaai.v32i1.11887.

[103] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.

[104] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[105] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[106] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[107] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*.

[108] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7553–7563.

[109] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2642–2651.

[110] F. Wang, L. Zhao, X. Li, X. Wang, and D. Tao, "Geometry-aware scene text detection with instance transformation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1381–1389.

[111] X. Wang, Y. Jiang, Z. Luo, C.-L. Liu, H. Choi, and S. Kim, "Arbitrary shape scene text detection with adaptive text region representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6442–6451.

[112] Y. Gao, Y. Chen, J. Wang, and H. Lu, "Reading scene text with attention convolutional sequence modeling," 2017, *arXiv:1709.04303*.

[113] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "AON: Towards arbitrarily-oriented text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5571–5579.

[114] S. K. Ghosh, E. Valveny, and A. D. Bagdanov, "Visual attention models for scene text recognition," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Nov. 2017, pp. 943–948.

[115] T. He, Z. Tian, W. Huang, C. Shen, Y. Qiao, and C. Sun, "An end-to-end TextSpotter with explicit alignment and attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5020–5029.

[116] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2315–2324.

[117] J. Ye, Z. Chen, J. Liu, and B. Du, "TextFuseNet: Scene text detection with richer fused features," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 516–522, doi: 10.24963/ijcai.2020/72.

[118] X. Liu, R. Zhang, Y. Zhou, Q. Jiang, Q. Song, N. Li, K. Zhou, L. Wang, D. Wang, M. Liao, M. Yang, X. Bai, B. Shi, D. Karatzas, S. Lu, and C. V. Jawahar, "ICDAR 2019 robust reading challenge on reading Chinese text on signboard," 2019, *arXiv:1912.09641*.

[119] C.-K. Ch'ng, C. S. Chan, and C.-L. Liu, "Total-text: Toward orientation robustness in scene text detection," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 23, no. 1, pp. 31–52, Mar. 2020, doi: 10.1007/s10032-019-00334-z.

[120] N. Nayef, Y. Patel, M. Busta, P. N. Chowdhury, D. Karatzas, W. Khlif, J. Matas, U. Pal, J.-C. Burie, C.-l. Liu, and J.-M. Ogier, "ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition-RRC-MLT-2019," 2019, *arXiv:1907.00945*.

[121] T. Jain, C. Lennan, Z. John, and D. Trans. (2019). *Imagededup*. [Online]. Available: https://github.com/idealo/imagededup

[122] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 970–983, May 2014.

[123] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 3304–3308.

[124] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *Int. J. Comput. Vis.*, vol. 116, no. 1, pp. 1–20, Jan. 2016.

[125] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao, "Scene text detection via holistic, multi-channel prediction," 2016, *arXiv:1606.09002*.

[126] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2014, pp. 3104–3112.

[127] F. Yin, Y.-C. Wu, X.-Y. Zhang, and C.-L. Liu, "Scene text recognition with sliding convolutional character models," 2017, *arXiv:1709.01727*.

[128] Z. Liu, Y. Li, F. Ren, W. L. Goh, and H. Yu, "SqueezedText: A real-time scene text recognition by binary convolutional encoder–decoder network," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7194–7201.

[129] S. Qin, A. Bissaco, M. Raptis, Y. Fujii, and Y. Xiao, "Towards unconstrained end-to-end text spotting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4703–4713.

[130] C. Luo, Q. Lin, Y. Liu, L. Jin, and C. Shen, "Separating content from style using adversarial learning for recognizing text in the wild," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 960–976, Apr. 2021.

[131] Y. Mou, "PlugNet: Degradation aware scene text recognition supervised by a pluggable super-resolution unit," in *Proc. ECCV*, 2020, pp. 158–174.

[132] W. Wang, "Scene text image super-resolution in the wild," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 650–666.

[133] J. Chen, B. Li, and X. Xue, "Scene text telescope: Text-focused scene image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12021–12030.

[134] M. Huang, Y. Liu, Z. Peng, C. Liu, D. Lin, S. Zhu, N. Yuan, K. Ding, and L. Jin, "SwinTextSpotter: Scene text spotting via better synergy between text detection and text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4583–4593.
[135] J. Ma, Z. Liang, and L. Zhang, "A text attention network for spatial deformation robust scene text image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5901–5910.

TIEN DO received the master's degree in computer science in 2016. He is currently pursuing the Ph.D. degree. He works as a Lecturer with the Faculty of Computer Science, University of Information Technology (UIT), which is part of Vietnam National University Ho Chi Minh City (VNU-HCM). He specializes in computer vision, namely in the areas of scene text recognition and object detection.

THUYEN TRAN received the bachelor's degree in computer science from the University of Information Technology in 2022. He is currently a Lecturer with the University of Information Technology. His primary area of interests include revolves around computer vision, with a specific focus on the fascinating field of optical character recognition (OCR).

THUA NGUYEN received the bachelor's degree in computer science from the University of Information Technology, VNU-HCM, in 2019. Currently, he works as a Research Scientist with the Multimedia Communications Laboratory (MMLab), University of Information Technology, VNU-HCM. His research interests include computer vision, with a primary focus on facial recognition and optical character recognition.

DUY-DINH LE (Member, IEEE) received the bachelor's and master's degrees from the University of Science, Ho Chi Minh City, Vietnam, in 1995 and 2001, respectively, and the Ph.D. degree from The Graduate University for Advanced Studies (SOKENDAI), Japan, in 2006. He was an Associate Professor at the National Institute of Informatics (NII), Japan, from 2013 to 2016. He is currently the Scientist and a Lecturer with the University of Information Technology, Vietnam. His research interests include semantic concept detection, video analysis and indexing, pattern recognition, machine learning, and data mining.

THANH DUC NGO received the Ph.D. degree from The Graduate University for Advanced Studies (SOKENDAI) in 2013. He has been a Lecturer with the Faculty of Computer Science, University of Information Technology (UIT), Vietnam National University Ho Chi Minh City (VNU-HCM), since 2014. His research interests include computer vision and multimedia content analysis.