Street View Text Recognition With Deep Learning for Urban Scene Understanding in Intelligent Transportation Systems

Chongsheng Zhang, *Member, IEEE*, Weiping Ding, *Senior Member, IEEE*, Guowen Peng, Feifei Fu, and Wei Wang

***Abstract*— Understanding the surrounding scenes is one of the fundamental tasks in intelligent transportation systems (ITS), especially in unpredictable driving scenes or in developing regions/cities without digital maps. Street view is the most common scene during driving. Since streets are often full of shops with signboards, scene text recognition over the shop sign images in street views is of great significance and utility to urban scene understanding in ITS. To advance research in this field, (1) we build *ShopSign*, which is a large-scale scene text dataset of Chinese shop signs in street views. It contains 25,770 natural scene images, and 267,049 text instances. The images in ShopSign were captured in different scenes, from downtown to developing regions, and across 8 provinces and 20 cities in China, using more than 50 different mobile phones. It is very sparse and imbalanced in nature. (2) we carry out a comprehensive empirical study on the performance of state-of-the-art DL based scene text reading algorithms on ShopSign and three other Chinese scene text datasets, which has not been addressed in the literature before. Through comparative analysis, we demonstrate that language has a critical influence on scene text detection. Moreover, by comparing the accuracy of four scene text recognition algorithms, we show that there is a very large room for further improvements in street view text recognition to fit real-world ITS applications.**

***Index Terms*— Street view, scene understanding, natural scene text recognition, Chinese photo OCR, digital mapping.**

I. Introduction

DIGITAL mapping plays an increasingly important role in intelligent transportation systems (ITS). It is the process by which a vast amount of data (satellite imagery, 3D aerial imagery, street view imagery, and road information) is compiled and formatted to produce a virtual map that gives accurate representations of a particular area, detailing road arteries and points of interests (POIs). However, in this process, POIs collection relies heavily upon manual or semi-automatic annotations, which consume a huge amount of labor and time. In recent years, as one of the key enabling technologies for autonomous driving, high definition (HD) maps have become a big industry and a major research focus [1], [2].

Besides HD maps, autonomous driving cars also use RGB camera images, LiDAR points and Radar points for multi-modal signal perception. A robust and reliable perception system is a prerequisite for driver-less cars to run safely in uncontrolled and complex driving environments [3], such as complex urban scenarios which involve persons, vehicles, traffic signs and road signs, and unforeseen circumstances, e.g., a child hidden in the blind areas suddenly runs into the roads. For safety considerations, it is always essential to understand the surrounding scenes during driving. For instance, the vehicles should slow down near the schools or accesses to pedestrian areas. However, in computer vision, it is often ambiguous to identify certain complex or confusing scenes, yet the texts contained in the scene images often provide critical information to understand such scenes [4]. Urban streets are often full of shops with signboards. Therefore, street view text recognition will be one of the enabling techniques for safe autonomous driving. In particular, in places without digital map and street view coverage, such as small towns in the mountain area or in the undeveloped countries or places, real-time street view text recognition will be very helpful for scene understanding during autonomous driving.

Recent advancements in perception for autonomous driving are driven by deep learning (DL), e.g., [5]–[7]. However, one of the main limitations preventing deep learning from being applied to a large set of street view text recognition and urban scene understanding is the lack of an extensive dataset with different scenes and a sufficient number of instances for each scene, since DL based methods are data-driven, that is, they need to consume large numbers of samples to achieve good performance [7], [8]. With algorithms/systems which are able to automatically extract the explicit textual contents (entity names and related information) from the signboard images, together with its corresponding position information [5], we will be able to realize the automatic

Manuscript received May 27, 2020; revised July 27, 2020; accepted August 13, 2020. The work of Weiping Ding was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20191445, and sponsored by Qing Lan Project of Jiangsu Province. The Associate Editor for this article was A. Jolfaei. *(Corresponding author: Weiping Ding.)*

Chongsheng Zhang, Guowen Peng, and Feifei Fu are with the Henan Provincial Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng 475001, China, and also with the School of Computer and Information Engineering, Henan University, Kaifeng 475001, China (e-mail: chongsheng.zhang@yahoo.com; gwpeng@henu.edu.cn; fff@henu.edu.cn).

Weiping Ding is with the School of Information Science and Technology, Nantong University, Nantong 226019, China (e-mail: dwp9988@163.com).

Wei Wang is with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China (e-mail: wangwei1@bjtu.edu.cn).

Digital Object Identifier 10.1109/TITS.2020.3017632

Fig. 1. Sample images of ShopSign.

(a) mirror (b) wooden (c) deformed (d) exposed (e) obscured

Fig. 2. Examples of the five categories of hard images in ShopSign.

collection of POIs, which is of great significance and utility in digital maps and scene understanding. Overall, street view signboard text recognition techniques will be beneficial for local geographical environment understanding in ITS; they will be also very helpful for precise positioning in autonomous driving scenarios, where the extracted signboard entities can be used as anchor points in the positioning process.

Robust reading from natural scene images that contain texts (a.k.a, scene text detection and recognition, or "Photo OCR") is an important research field that has many real-world applications, including digital mapping, autonomous driving, scene understanding, location based services (LBS), content-based image retrieval, etc. Hence, a significant amount of research effort has been invested in this area over the past decade, especially with the rise of deep learning. In the literature, CTPN [9], DeepDirect [10], EAST [11], and TextBoxes++ [12] are representative DL based methods for scene text detection, while CRNN [13], SlidingCNN [14], and ASTER [15] are well-known algorithms for DL based scene text recognition. However, existing techniques focus on Latin scripts, whereas little effort has been invested in Chinese scene text detection and recognition. This is not only because of its inherent complexity and difficulty, but also due to the scarcity of large-scale well-annotated Chinese scene text datasets, since DL based techniques are data-driven, data-eager. In this work, we mainly deal with Chinese Photo OCR over street view shop sign images, which is more challenging and difficult than English Photo OCR.

Reading shop signs in street views, which is a specific sub-field (application) of scene text detection and recognition, is of great significance to digital mapping and urban scene understanding. The names and locations of shops along the streets are a vital source of POI data which is required by digital maps; however, these POIs are often manually (or semi-automatically) annotated by workers. To promote research in recognizing/understanding Chinese shop signs in street views, we build a specialized and challenging dataset which consists of shop sign images along the streets in China. This dataset is hereafter referred to as "ShopSign" [16]. It contains 25,770 images collected from more than 20 cities, using 50 different smart phones. These images exhibit a wide variety of scales, orientations, lighting conditions. Moreover, we characterize the difficulty of ShopSign by specifying five categories of "hard" images, which contain mirror, exposed, obscured, wooden, or deformed texts. The images in ShopSign have been manually annotated in "text-line" manner by 10 research assistants. In Figures 1 and 2, we showcase a few representative images and hard images in ShopSign.

Moreover, we make a comprehensive evaluation of existing DL based scene text detection and recognition algorithms on ShopSign and another three Chinese scene text datasets, which has not been studied in the literature so far. We note that, by "Chinese images" ("Chinese datasets"), we denote the natural scene images (datasets) having Chinese as the main script for the texts contained in the images. We also clarify that, "Chinese Photo OCR" is a subfield of "Chinese

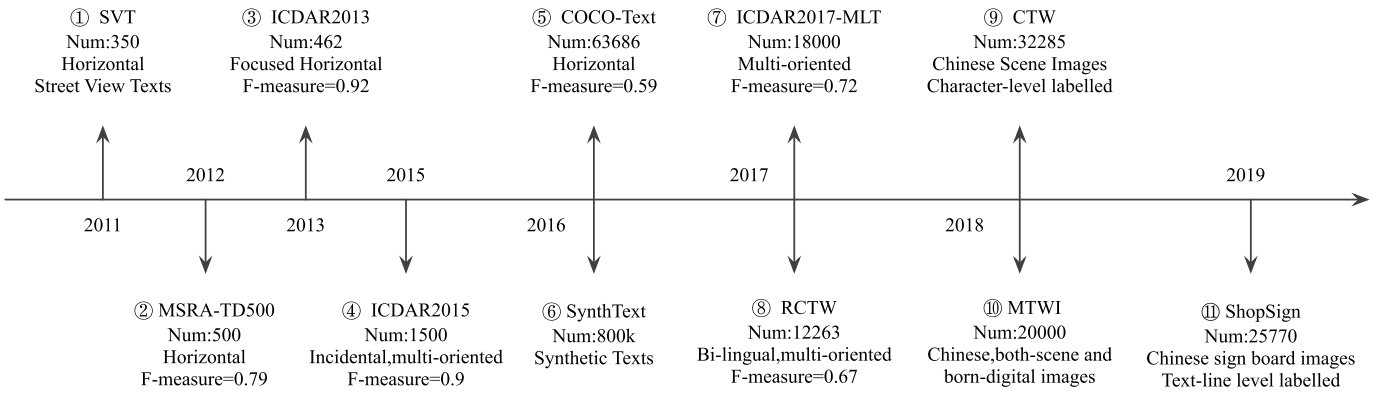

Fig. 3. The development of scene text datasets along time.

OCR", the former focuses solely on Chinese texts detection and recognition from natural scene images, while the latter also includes other subjects, such as document image analysis.

For Chinese scene text localization from the natural scene images, we use TextBoxes++ and EAST, which are state-of-the-art methods for DL based scene text detection, to train text detection models from English scene text datasets and test their performance on the Chinese natural scene images, and vice versa. Our experiments show that models trained from English scene text datasets using EAST and TextBoxes++ achieve low recall and precision results on the Chinese scene text datasets. This demonstrates that script/language of the scene images substantially influences text detection, hence the need of Chinese natural scene image sets for Chinese scene text detection and urban scene understanding.

For Chinese scene text recognition, we employ CRNN, SlidingCNN and ASTER, and evaluate their recognition performance on ShopSign and the other Chinese scene text datasets. We also introduce SliceCNN, which is an extension of SlidingCNN. The proposed SliceCNN method is based on our intuition that, even though a full character is split into a few vertical pieces/stripes, with CTC [17] it is still possible to infer the full character from the sequential predictions for these stripes. We conduct comprehensive empirical studies to compare the recognition performance of these algorithms. We show that ASTER is the best performer for Chinese scene text recognition, but there is still a great room for further improvements.

The main contributions of this work can be summarized as follows:

1) We build ShopSign, which is a large-scale street view shop sign dataset. It will have important applications in digital mapping and scene understanding in intelligent transportation systems.
2) We investigate street view text recognition methods on ShopSign, and conduct a comprehensive empirical evaluation on the performance of existing DL-based scene text reading methods on this dataset.
3) We show that existing DL-based scene text recognition algorithms achieve low recognition accuracy on ShopSign. This reveals that street view text recognition is a highly challenging research direction with lots of potential.
4) We attest that DL-based street view text detection is significantly language/script dependent, which conflicts with the common perception.

The remainder of this article is organized as follows. In Section II, we introduce related work in scene text reading. Next, in Section III, we introduce the problem definition, the ShopSign dataset, and techniques for street view text recognition. In Section IV, we design experiments and analyze the results of existing DL-based scene text reading algorithms on ShopSign. Then, in Section V, we further check the influence of script/language on street view text detection. In Section VI, we address the importance and benefits of street view text recognition techniques to scene understanding and ITS, then summarize our experimental findings from this work. Finally, we conclude the paper in Section VII.

II. Related Work

In the past few years, with the development and rise of deep learning, a prevalent trend in scene text detection and recognition has been the shift from conventional methods (e.g., [18]) to DL based methods (e.g., [11], [19]. A rich body of literature has been published in recent years, readers may refer to [20]–[23] for comprehensive surveys. In this section, we will briefly summarize existing scene text datasets and DL based scene text detection and recognition algorithms.

A. Scene Text Datasets

From ICDAR 2003 [24], several benchmark datasets have been constructed for English scene text detection and recognition. Figure 3 depicts the development of scene text datasets along time. Among them, the most widely used datasets are ICDAR 2013 [25] and 2015 [26]. ICDAR 2013 [25] is dataset of images having focused texts, while ICDAR 2015 [26] is an image dataset with incidental texts. Besides, MSRA-TD500 [27] is also a frequently used dataset for scene text detection.

The MLT dataset [28] released in ICDAR 2017 (and continued in ICDAR 2019) is a very interesting multi-lingual dataset

with 9 (10 in ICDAR 2019) different languages (each language has 2,000 images). It is used mainly in multi-lingual scene text detection [29].

For English scene text recognition, SVT [30], IC13 [24], [25], and IIIT5k [31] are the commonly adopted datasets, whereas RCTW [32], MTWI [33], CTW [34], and ShopSign [16] are the currently available datasets for Chinese scene text recognition (note that they can be used in Chinese scene text detection as well).

RCTW (2017) [32] is the first scene text dataset that contains both English and Chinese scripts(we note that, although some images in MSRA-TD500 [27] also contain Chinese scripts, this dataset only provided coordinate annotations of the text instances, but without the ground-truth texts.). The images were captured outdoors or through screen shots. MTWI (2018) [33] is a competition dataset of a well-known international conference (ICPR 2018). This dataset was provided by a large e-commerce company (Alibaba), and many of the images in this dataset are born-digital, because most of them are promotional pictures. CTW (2018) [34] and ShopSign (2019) [16] are two Chinese natural scene datasets which can be used for evaluating both scene text detection and recognition algorithms. The images of CTW were collected using street view collection vehicles equipped with unified cameras (with high resolution) and annotated at character level, while the images in ShopSign were harvested by 40 research assistants using 50 different smart phones and cameras, and labeled at text-line level.

Recently, there are two large-scale Chinese scene text datasets that are publicly available, which are the ICDAR-2019 LSVT[1] and ReCTS[2] datasets. LSVT consists of 50,000 fully annotated images and 400,000 weakly labeled ones. However, most of the text instances in LSVT are (near) horizontal ones. ReCTS offers both text-line and character-level annotations. Moreover, it also provides one or more ground truths for each test image if needed, since the definition of "text-lines" is sometimes ambiguous. The number of images in ReCTS is close to ShopSign, and both datasets are signboards oriented. However, the images in ReCTS merely contain the signboards, but not the background building and street views. In comparison, most images in ShopSign contain both the signboards and natural scene backgrounds.

In this work, ICDAR 2013, ICDAR 2015 and SynthText [35] will be adopted for training English scene text detection models, whereas the four Chinese benchmark datasets RCTW, MTWI, CTW, and ShopSign will be used for Chinese scene text detection and recognition.

B. DL Based Scene Text Detection

Most of existing DL based scene text detection algorithms graft object detection/image segmentation techniques in computer vision/deep learning to scene text reading application scenarios. In particular, Faster R-CNN [36] is the widely adopted technique used in scene text detection [9], [37], [38], owing to its excellent performance in object detection from images. The models are often pre-trained on ImageNet which is a large repository with millions of annotated images. Besides, recent object detection algorithm YOLO [39] has also been adapted to scene text detection. TextBoxes++ [12], [40], which is one of the latest advances in text detection, adopts the Single Shot Detector (SSD) [41] network structure, which is a very well-known technique for object detection. Table I gives a short summary of the computer vision techniques used by existing scene text detection algorithms.

TABLE I

TAXONOMY OF DL BASED SCENE TEXT DETECTION ALGORITHMS BASED ON THE TYPES OF NEURAL NETWORK STRUCTURES ADOPTED

Method \ Category	Object Detection Based	Segmentation Based
CTPN [9]	Faster R-CNN	
[37, 38]	Faster R-CNN	
SegLink [43]	SSD	
TextBoxes++ [40, 12]	SSD	
EAST [11]		FCN
DeepDirect [10]		FCN

Fully Convolutional Networks (FCN) [42] is a classical DL based image segmentation algorithm, which has been adapted to scene text detection scenarios as well. Both EAST [11] and DeepDirect [10] use FCN for text-line detection.

Among the recent advances in DL based scene text detection, CTPN [9], EAST [11], DeepDirect [10], and TextBoxes ++ [12], [40] are representing methods. CTPN localizes a sequence of fine-scale text proposals by densely sliding a small window in the convolutional feature maps; next, an in-network recurrence mechanism is used to connect the sequential text proposals. It achieves outstanding performance in horizontal text-line detection.

Both EAST [10] and DeepDirect [10] regress text boxes directly from the dense segmentation maps generated by FCN. DeepDirect utilizes FCN feature maps for both pixel-wise classification of the text/non-text pixels, and text boxes regression to determine the coordinates of the quadrilateral text boundaries. In comparison, EAST devises an FCN-based pipeline that directly produces word or text-line level predictions (in terms of quadrilateral or rectangles), which are then merged using Non-Maximum Suppression (NMS). TextBoxes [40] employs the SSD network structure but adopts long convolutional filers to adapt to the long text-line detection scenario. TextBoxes++ [12] extends TextBoxes from horizontal to multi-oriented scene text detection, through regressing the coordinates of the quadrilaterals.

CRAFT, recently proposed in [44], is a representative method for character-level scene text detection that obtained state-of-the-art performance. It first localizes the individual character regions using a Gaussian heatmap on every character's center, then links them into a text instance. To compensate for the lack of character-level annotations, they propose a weakly-supervised learning framework to estimate the character-level ground truths in existing datasets.

Arbitrary shape scene text detection. In [45], the authors propose a relational reasoning graph network for arbitrary shape text detection. Every text instance is divided into a

[1]https://rrc.cvc.uab.es/?ch=16
[2]https://rrc.cvc.uab.es/?ch=12

sequence of small rectangular components, and their geometry attributes will be used by a local graph model to roughly establish linkages between different components. For further reasoning and deducing the likelihood of linkages between the component and its neighbors, a graph-based network is adopted to perform deep relational reasoning on the local graphs. This is the first attempt to use graph convolutional network for arbitrary shape text detection, which achieved the state-of-the-art performance on both multi-oriented and arbitrary shape scene text datasets.

In [46], the authors present ContourNet, which is able to effectively suppress false positive text proposals. By using a set of pre-defined points instead of the 4-d vector for the proposal representation, ContourNet is also adaptive to arbitrary shape text regions.

ABCNet [47] is a real-time scene text detection method, which adaptively fits oriented or curved text by a parameterized Bezier curve. It achieved state-of-the-art performance and is significantly faster than previous methods.

C. DL Based Scene Text Recognition

Only in recent years, people have started investigating DL based scene text recognition, where CRNN, SlidingCNN, and ASTER are typical methods. All the DL based scene text recognition methods utilize CNNs for feature extraction from scene images. The main differences among them lie in the transcription/decoding step, which transforms the predictions on the Convolutional feature maps into a sequence of labels, and the image pre-processing step before CNN. CTC [17] and attentional mechanisms are the two major types of decoding methods, and the transcription layers in most of the existing scene text recognition methods fall into these two categories.

1) CTC Based Transcription: CRNN [13] and SlidingCNN [14] are well-known CTC based scene text recognition algorithms. CRNN [13] is a novel neural network architecture that combines CNN and RNN for scene text recognition. It first uses CNN to extract a feature sequence from an input image, then applies RNN (BLSTM) to make prediction for each frame of the feature sequence. Finally, a transcription layer (CTC) [17] is adopted to translate the per-frame predictions into a label sequence. CRNN is end-to-end trainable, since CNN and RNN are jointly trained with only one loss function.

SlidingCNN [14] aims at recognizing texts from text-line crops. It comprises a sliding window layer, a classification layer and a transcription (CTC) layer. From each window along the text-line crop, this method applies CNN to extract the convolutional features, upon which per-window predictions are made. These sequential per-window predictions are finally fed into the CTC layer to be transcripted into a sequence of character labels. There are two differences between SlidingCNN and CRNN. First, SlidingCNN eliminates the RNN (BLSTM) layer. Second, CRNN extracts the convolutional features from the whole image, whereas SlidingCNN obtains such features from each sliding window along the text-line, and makes per-window predictions.

Besides, STAR-Net [48] also relies on CTC for decoding. Before extracting the convolutional feature maps, it first uses a spatial transformer to convert the distorted texts into rectified textual regions. Experimental results on five public benchmark datasets show that STAR-Net outperforms other methods on scene texts with considerable distortions, while still achieving a performance comparable to other methods on scene texts with little distortions. Since its source codes are not publicly available, we do not include it in our experiments.

2) Attention Based Transcription: ASTER [15] is an advanced DL based method that uses an attentional mechanism for scene text recognition. It comprises a rectification network and a recognition network: the former adaptively transforms an input image into a new one with the irregular/perspective texts being rectified, while the latter predicts a character sequence from the rectified image using an attentional sequence-to-sequence transcription approach. Overall, ASTER is similar to CRNN, except the following two differences. First, ASTER adds an additional rectification network (image pre-processing step) to spatially transform the irregular/perspective texts of a crop. Second, CRNN uses CTC for transcription, whereas ASTER adopts an attentional sequence-to-sequence model as the decoder. In their implementation, the authors use a 45-layer ResNet network structure for convolutional feature extraction.

SqueezedText [49] is a real-time scene text recognition method which consists of a binary convolutional network that extracts binary format features, and a bi-directional RNN for character level classification. It achieves very fast processing speed and comparable recognition accuracy.

In [50], the authors propose a semantics enhanced encoder-decoder framework (SEED) for scene text recognition, which can predict semantic information using language models and robustly recognize low-quality scene texts. In [51], the authors propose SCATTER which receives features from two different layers in the neural network, namely, visual features from a CNN backbone and contextual features computed by cascaded BiLSTM layers (named as "deep BiLSTM encoders") in the training process.

Arbitrarily-oriented/irregular text recognition is a recent focus of the community [52]–[56]. Show-Attend-and-Read [54] is an irregular text recognition algorithm that is composed of a convolutional layer, an LSTM-based encoder-decoder framework, and a tailored 2-dimensional attention module for handling the complicated spatial layout of irregular texts. It achieves state-of-the-art performance in irregular scene text recognition and favorable performance on regular texts.

TextScanner [57] is a segmentation-based dual-branch (which are class and geometry branches) framework for scene text recognition. It segments characters one by one and ensures that characters are read in right order and separated properly. In [58], the authors propose an effective convolutional character networks (referred as *CharNet*) for joint character detection and recognition. Since *CharNet* requires character-level scene text annotations, the authors propose a heuristic iterative character detection approach to tackle this problem. Character Attention Fully Convolutional Network (CA-FCN) [59] aims at recognizing arbitrarily oriented texts. It first uses FCN to predict characters at pixel level (we note that character-level annotations are needed by CA-FCN), then a character attention module to highlight the foreground

characters and weaken the background. It outperforms existing methods with a large margin on irregular datasets, and still achieves excellent performance on regular datasets.

In [60], the authors introduced a unified four-stage framework for DL based scene text recognition, which consists of transformation (rectification), feature extraction (various network structures to extract CNNs), sequence modeling (e.g., BiLSTM) and prediction(transcriptions using CTC or attentional mechanism). Using this framework, the authors conducted an extensive evaluation of previously proposed scene text recognition modules. However, they only used English scene text datasets in their study.

In [4], Bai *et al.* investigated the integration of visual representation and textual contents for fine-grained image classification with Convolutional neural networks, which significantly outperforms classification with only visual representation. It also improves the content-based image retrieval performance by a large margin.

In [61], the authors proposed an efficient approach to attacking (i.e., automatically recognizing) variable-length Chinese character CAPTCHAs with noises. However, the backgrounds for text-based CAPTCHAs are usually gray-scale or relatively clean, which is very different from street view scene images. In [62], the authors propose a generic attack method against both CTC-based and attention-based scene text recognition (STR) models. They further use their adversarial examples to attack commercial STR system and corrupt the predictions in many cases.

In [63], the authors propose an end-to-end network for text editing task, which can replace the original text in the scene text images while still maintaining the original style.

The authors in [64] propose a data augmentation method for text images that contain multiple characters, as well as a framework that jointly optimizes the data augmentation and the recognition modules for effective STR model training.

Although currently sequence-to-sequence learning is the mainstream approach for scene text recognition and has obtained state-of-the-art performance, such methods have performance bottlenecks, because they need to consume huge amounts of labeled images (especially for Chinese Photo OCR) to obtain satisfactory performance, which are very hard to collect and annotate. Character-level scene text localization and recognition (e.g., [44], [58], [57]) is less dependent on the data scale, which will be a promising direction for future Photo OCR research. But this will need layout analysis technique, which is still in the early research stage on scene text images.

III. Street View Text Recognition

A. Problem Statement

Driving scene understanding is a key module in intelligent transportation systems for autonomous driving. The performance of visual scene recognition tasks has been significantly boosted by recent advances of deep learning algorithms [65]. However, these advances are mainly about object detection and semantic segmentation of the traffic participants on the road [66], [67], while little attention has been paid to the surrounding scene understanding in the ITS field.

The main innovation of this work lies in the collection and utilization of street view shop sign images for surrounding scene understanding in ITS. Street view text recognition techniques to be developed upon our ShopSign street scene text dataset have three important effects: (1) surrounding scene understanding in intelligent transportation systems, which further promotes the understanding of the driving scenes so that autonomous vehicles can effectively adapt to different street scenes and reduce accidents. (2) automatic POIs collection (the entity names of the shops and its spatial location), which will significantly enrich the contents of digital maps. This in turn will be very helpful for autonomous driving and intelligent transportation systems. (3) precise positioning, where the large amount of extracted shop entities can be used as important anchor points in the positioning process. All these three functions are very useful in ITS and autonomous driving.

B. Building the ShopSign Dataset

1) Dataset Collection and Annotation: In developed countries such as USA, Italy and France, there are very limited number of characters in the signboards, and their sizes are usually small. Owing to the differences in language, culture, and history, Chinese shop signs have distinctive features. Even inside China, there is a big diversity in the materials and styles of the shop signs across different regions. For instance, shops in major cities usually adopt fiber-reinforced plastic and neon sign boards; but in the suburb or developing regions, economic wooden and outdoor inkjet and acrylic shop signs are very common. The styles of the shop signs also vary in different provinces, e.g., shop signs in Inner Mongolia and Xinjiang provinces are different from Shanghai.

Building a large-scale dataset of Chinese shop signs is a fundamental yet critical task that needs enormous manual collection and annotation effort. We obtained help from 40 volunteers of our institution and spent more than 2 years in collecting the shop sign images in 20 different cities/regions of China, and in the annotation of these images. A total of 50 different cameras and smart phones were used in the collection and many of the images carry GPS locations. Two faculty members and ten graduate students were involved in the annotation of these images (in text-line manner using quadrilaterals). Finally, the ShopSign dataset we build contains 25,770 well annotated images of Chinese shop signs. In Figures 1 and 2 showcase a few sample images in ShopSign.[3]

2) Dataset Statistics: In this subsection, we provide more detailed statistics of ShopSign, including: (1) the resolution (scale) of text instances; (2) the orientation of the text instances; (3) the language of the text instance; (4) the number of hard images in the five special categories. Summarized in Table II, ShopSign contains 25,770 Chinese shop sign images and 267,049 text instances. The total number of unique Chinese characters is 4,072, with 626,280 occurrences in total. Figure 5 depicts the ratios of characters with different occurrence ranges.

[3]More sample images of ShopSign are publicly available at: https://github.com/chongshengzhang/shopsign.

(a) Pair Images (I)-1 (b) Pair Images (I)-2 (c) Pair images (II)-1 (d) Pair images (II)-2

Fig. 4. Two examples of paired images in ShopSign.

TABLE II
BASIC STATISTICS OF SHOPSIGN

Item	Number
Total Number of Images	**25,770**
Training images	20,738
Testing images	5,032
Chinese Characters	**626,280**
Unique Chinese Characters	**4,072**
Average number of Text Instances Per Image	10.4
Orientation of Text Instances	
Total number	**267,049**
Horizontal	186,524
Vertical	48,744
Multi-directional	31,781
Script of Text Instances	
Chinese	**157,261**
English	21,517
Chinese-English mixed	4,624
others	83,647
Resolution of Text Instances	
Average Aspect Ratio (Horizontal)	3.10
Average Aspect Ratio (Vertical)	2.16
Average width	529 pixels
Average height	239 pixels
Maximum width	5,308 pixels
Maximum height	5,248 pixels
Minimum width	4 pixels
Minimum height	4 pixels

Here, the orientation of a text instance is derived via the angle between the text instance and the horizontal/vertical axis of the image where the text instance resides. On the transverse direction, there are 210,987 text instances, including 50,075 horizontal ones (orientation = 0 degree), and 136,449 instances with orientation less than 10 degrees, and the other 24,463 instances have orientations larger than 10 degrees. On the longitudinal direction (when the height of a text instance is larger than its width), there are 5,957, 42,787, and 7,339 text instances with orientation angles being 0, <=10 and >10, respectively. In total, there are 186,524 horizontal text instances (orientation angle <=10), 48,744 vertical text instances, and 31,781 multi-directional text instances; their corresponding ratios are 69.85%, 18.25% and 11.90%, respectively.

Concerning the scripts of the text instances, there are 157,261 text instances with Chinese as its only script (referred to as Chinese instances for short), 21,517 English instances, and 4,624 Chinese and English mixed text instances. The rest 83,647 text instances contain purely numerals or the "#" symbols (which denotes unrecognizable texts).

3) Dataset Characteristics: Overall, ShopSign has the following characteristics:

1) **Large-scale and specialized in signboards**. It comprises 25,770 natural scene images, all of which are Chinese signboards in street views.
2) **Night images**. It includes near 4,000 night images (captured in the night). In such images, signboards are very remarkable and the rest background areas are dark. Such night images rarely exist in other datasets.
3) **Special categories of hard images**. ShopSign consists of 5 special categories of hard images (the total number is 1,746), which are *mirror*, *wooden*, *deformed*, *exposed* and *obscure*, as depicted in Figure 2. Their numbers of images are 106, 823, 152, 85 and 580, respectively. Text detection and recognition over such hard images should be more challenging than ordinary natural scene images.
4) **Sparsity and class imbalance**. The number of unique characters with 500 or more occurrences is 333 (8.2% in ratio), as can be seen from Figure 5, but the sum occurrences of these characters occupies 64.7% of the total character occurrences. In comparison, the ratio of characters with 100 or less occurrences is 74.7%, yet their total number of occurrences is only 9.5%. Furthermore, 537 characters only have 1 occurrence, and 1,687 characters (41.4% in ratio) have 10 or less occurrences. Hence, the distribution of Chinese character occurrences in ShopSign is highly skewed. By sparsity, we mean that most characters occur in only a few signboard images.
5) **Pair images**. Our dataset contains 2,516 pairs of images. In each pair of images, the same signboard was shot twice, from both frontal and tilted perspectives. Pair images greatly facilitate the evaluation/comparison of an algorithm's performance on horizontal and multi-oriented text detection. In Figure 4, we show two examples of pair images in ShopSign.

C. Street View Text Recognition Techniques

For street view text recognition in scene understanding, we carry out an extensive empirical study on the performance of state-of-the-art scene text detection and recognition algorithms on ShopSign. We use EAST and TextBoxes++ for scene text detection, and CRNN, ASTER, and SlidingCNN for scene text recognition. As a minor contribution, we also propose the SliceCNN algorithm for shop sign text recognition.

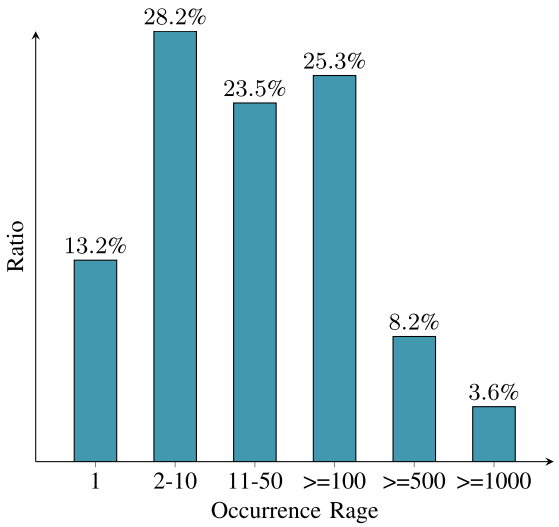

Fig. 5. Ratios of characters with different occurrence ranges.

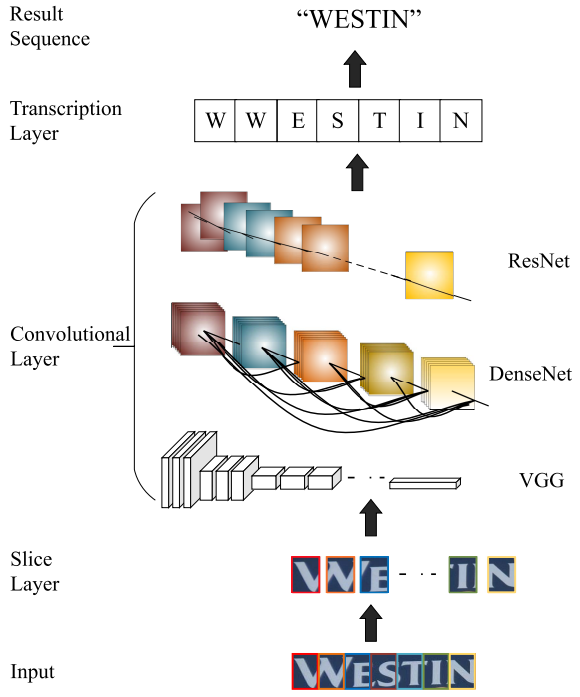

Fig. 6. The framework of the SliceCNN method.

Figure 6 depicts the overall framework of the proposed SliceCNN algorithm [14]. The main difference from SlidingCNN lies in the slice layer, which replaces the sliding window component of SlidingCNN. In SliceCNN, the text-line crop is equally cut into several vertical stripes/slices. For each slice, we use CNN to extract the features and make per-slice predictions. The outputs of these slices are finally fed into the CTC layer and the loss is minimized using gradient descents. The other procedures are the same as SlidingCNN.

SlidingCNN needs to shift a window (of a character size) along the text-line crop, then extracts the convolutional features from each window and makes per-window predictions, which are finally trascripted through CTC. However, even when a character is split into several adjacent stripes/slices, with CTC it is still possible to infer the corresponding character. Moreover, the window size and stride parameters may be hard to be determined. So we propose SliceCNN, which removes the window and saccades constrains from SlidingCNN. It provides an alternative for scene text recognition. The only parameter in SlideCNN is the stripe width. In our experiments, we fix it to 8 pixels.[4]

Algorithm 1 The Training Process of SliceCNN

Input: text-line instances (training) $Train$,
ground truth annotations GT.
Output: text recognition model

```
iter = 1, maxIter = 600,000;
batchSize = 64;
Stride = 8;
while iter ≤ maxIter do
    k = 1;
    loss = 0;
    while k ≤ batchSize do
        read next text-line instance t;
        Slices[] = divide t into slices with Stride;
        ns = length(t)/Stride;
        for i = 1; i ≤ ns; i++ do
            Prob[i] = CNN(Slices[i]);
        loss+= ctc_loss(Prob, ns, GT[t]);
        k ++;
    update the network model with loss;
    iter ++;
```

In Algorithm 1, we present the training process of SliceCNN. For each image t in $Train$, we first cut it into equal-width slices. For each slice $Slice[i]$ of t, we use CNN (convolutional neural networks) to extract its deep visual features and make predictions, which are saved in $Prob[i]$. Let the total number of classes in $Train$ be n, and the number of slices on t be ns, then $Prob[i]$ is an array of size n, denoting the probabilities that the (incomplete) character in $Slice[i]$ belongs to each class. Finally, we feed the $n \times ns$ matrix $Prob$ and the ground truth annotation GT[t] for t to the ctc_loss function and obtain the loss. The main innovation of our algorithm lies in that, even if we cut each image into several slices and predict the class probabilities for each (incomplete) slice, with CTC we can still infer the corresponding character sequence of each image.

For EAST, TextBoxes++, CRNN, and ASTER, we use the original implementations of the authors [11]–[13], [15]; our parameter settings are also consistent with them. For SlidingCNN, we use our own implementations, since its source codes are not publicly available. For ASTER, by default it requires each text-line crop to be resized into 256×64 (width, height) pixels, whereas the other text recognition algorithms require all the crops to be resized to 256×32. The window size and stride of SlidingCNN are set to 32×32 and 4 pixels,

[4]We also tested SliceCNN with the stripe parameter being 4 pixels, but observed little performance increase.

respectively. ASTER adopts the 45-layers RestNet deep neural network structure. In our implementations for SlidingCNN and SliceCNN, we use the ResNet-101 and DenseNet-201 network structures.

IV. Experiments on ShopSign

A. Setup

RCTW [32] has 8,034 and 4,229 images for training and testing. Because the ground-truth of its testing data is unavailable, we only use its training data in our experiments, for which we randomly pick 6,000 images as our training data, and the rest 2,034 images are kept as the testing data. They contain 35,120 and 8,800 text-lines, respectively. MTWI [33] has 10,000 images for training and the same amount for testing. Also due to the unavailability of the ground-truth in its testing data, we only use its training data in our experiments. In our split, 8,000 images are randomly sampled as training data and the rest 2,000 are kept for testing; they contain 120,740 and 21,010 text-lines, respectively. Originally, CTW [34] contains 25,887 images in its training data. Since the labels of its testing data are also unavailable, we only use its training data, which is further split into training and testing parts in our experiments. In our split, we randomly select 5,032 images as test set to be consistent with ShopSign, while the other images are used as training set. We remove the sentences (text-lines) that are smaller than 1KB in size because they are too vague to be recognized. After preprocessing, we obtain 69,156 text-lines as training data and 6,568 text-lines as testing data.

For ShopSign, we first split it into training (Train1) and testing sets (Test1), which contain 20,738 and 5,032 images, respectively. This split is hereafter referred to as *Split-1*. The testing set comprises 2,516 pairs of images, which can be used to compare and reveal the ability of an algorithm in detecting horizontal and multi-oriented scene texts. Train1 and Test1 will also be used for assessing the performance of text recognition algorithms. The collection of cropped text-lines from Train1 (since they have annotations) will be used for training text recognition models, while the set of cropped text-lines from Test1 will be used for testing their recognition performance. Next, in order to validate/verify whether it is dataset split that leads to the evaluation performance to be reported, we re-split ShopSign (hereafter referred to as *Split-2*) and re-run the experiments with more scene text detection algorithms. In *Split-2*, we randomly pick half of the 2516 pairs (which is the test set of *Split-1*) and swap them for the same number of images from the training set of *Split-1*. Thus, the sizes of the training and testing sets in *Split-2* remain the same as *Split-1*. Third, for specific evaluation of text detection performance on the five "hard" categories of images, ShopSign is re-split (denoted as *Split-3*) into new training (Train3) and testing sets (Test3). Test3 comprises half of the images from each of the five "hard" categories, whereas all the other images of ShopSign are used as the new training set, i.e., Train3. In short, Train1 and Test1 (*Split-1*) are used for evaluating both the text detection and recognition algorithms, whereas Train3 and Test3 (*Split-3*) are specially designed for assessing the performance of text detection algorithms on the five "hard" categories of images. Train2 and Test2 (*Split-2*) are used to further validate the performance of state-of-the-art scene text detection algorithms on ShopSign and reveal the challenges.

TABLE III

Text Detection Performance of EAST, TB++ and CTPN on Split-1

Datasets	Methods	Horizontal			Multi_Oriented		
		R	P	H	R	P	H
RCTW	EAST	0.532	0.371	0.437	0.505	0.412	0.454
	TB++	0.401	0.516	0.451	0.380	0.432	0.405
	CTPN	0.442	0.446	0.444	0.373	0.407	0.389
MTWI	EAST	0.360	0.250	0.295	0.319	0.274	0.294
	TB++	0.328	0.392	0.357	0.306	0.34	0.322
	CTPN	0.380	0.490	0.428	0.336	0.480	0.395
CTW	EAST	0.241	0.261	0.250	0.215	0.279	0.243
	TB++	0.346	0.084	0.135	0.313	0.075	0.121
	CTPN	0.182	0.371	0.244	0.161	0.379	0.226
ShopSign	EAST	0.584	0.364	0.448	0.579	0.410	0.480
	TB++	0.471	0.501	0.486	0.479	0.476	0.478
	CTPN	0.535	0.566	0.550	0.444	0.518	0.478

Model training on each dataset will finish in 20 epochs. To be consistent, in our experiments we will use the Intersection over Union (IoU, threshold $= 0.5$) criteria for evaluating the performance of the scene text detection algorithms, and Edit Distance[5] for assessing the accuracy results of text recognition methods.

B. Scene Text Detection on ShopSign

1) Baseline Scene Text Detection Experiments: We first present the performance of major scene text detection algorithms on ShopSign, using Train1 and Test1. Test1 contains 2,516 pairs of images, with each pair of images containing both horizontal and multi-oriented text instances, which can more comprehensively compare the relative performance of scene text detection models on horizontal and multi-oriented text instances. In Table III, R, P, and H denote the recall, precision and F-score results, respectively. We observe that, EAST achieves the best text detection performance on ShopSign, with a recall of 57.9%-58.4%. We also notice that CTPN obtains better text detection results than TextBoxes++ (TB++) on the horizontal test set of ShopSign, but the latter outperforms the former in multi-oriented scene texts.

In Figures 7 and 8 (which are based on Table III), we show the relative difficulty of ShopSign with respect to RCTW/MTWI/CTW on scene text detection, using EAST, TextBoxes++ and CTPN. We use their official training data to train the corresponding text detection models, then test them on Test1 (*Split-1*) of ShopSign. We see that, text detection models trained on CTW and MTWI by all the three algorithms only obtain a recall of 16.1%-38% on ShopSign, whereas EAST trained on RCTW obtains the best recall result on ShopSign, which is between 50.5% and 53.2%. Such performance might also be partially caused by the relatively low generalization capability of these algorithms/models. From Figure 8, it is also interesting to observe that, when training on

[5]https://en.wikipedia.org/wiki/Edit_distance

TABLE IV
EXTENSIVE TEXT DETECTION EXPERIMENTS ON SHOPSIGN, USING EAST, TB++, CTPN, TEXTSNAKE, DERPN AND TEXTFIELD

Datasets	Training	MSRA-TD500			MLT			RCTW			MTWI			ShopSign		
		R	P	H	R	P	H	R	P	H	R	P	H	R	P	H
EAST	alone	0.539	0.559	0.549	0.533	0.666	0.592	0.443	0.538	0.486	0.437	0.778	0.559	0.554	0.457	0.388
	+ ShopSign	0.575	0.534	0.554	0.537	0.671	0.596	0.449	0.572	0.503	0.435	0.776	0.558	-	-	-
TB++	alone	0.607	0.744	0.668	0.325	0.687	0.441	0.313	0.607	0.413	0.365	0.668	0.472	0.499	0.497	0.496
	+ ShopSign	0.591	0.616	0.603	0.315	0.652	0.425	0.319	0.596	0.415	0.368	0.595	0.455	-	-	-
CTPN	alone	0.444	0.413	0.428	0.155	0.333	0.211	0.328	0.423	0.373	0.263	0.469	0.337	0.31	0.316	0.323
	+ ShopSign	0.488	0.504	0.496	0.15	0.312	0.203	0.32	0.41	0.36	0.276	0.489	0.353	-	-	-
TextSnake	alone	0.499	0.428	0.461	0.211	0.419	0.281	0.307	0.536	0.390	0.382	0.554	0.452	0.305	0.286	0.269
	+ ShopSign	0.562	0.586	0.574	0.28	0.489	0.356	0.31	0.561	0.400	0.37	0.513	0.43	-	-	-
DeRPN	alone	0.682	0.22	0.333	0.542	0.213	0.306	0.653	0.127	0.213	0.6	0.323	0.42	0.751	0.411	0.283
	+ ShopSign	0.733	0.253	0.376	0.45	0.106	0.172	0.655	0.112	0.191	0.474	0.154	0.233	-	-	-
TextField	alone	0.57	0.671	0.616	0.391	0.705	0.503	0.191	0.621	0.292	0.27	0.721	0.393	0.649	0.35	0.239
	+ ShopSign	0.504	0.678	0.578	0.361	0.691	0.474	0.196	0.652	0.302	0.243	0.721	0.364	-	-	-

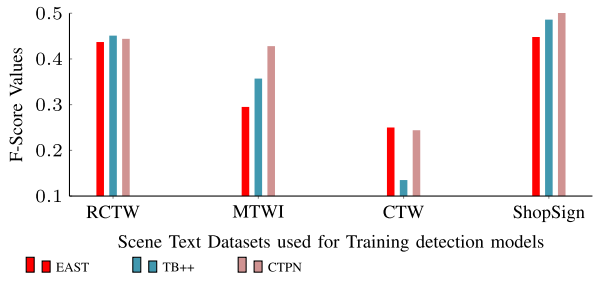

Fig. 7. F-scores of street view text detection algorithms EAST, TB++, CTPN trained on different datasets and tested on ShopSign (Horizontal).

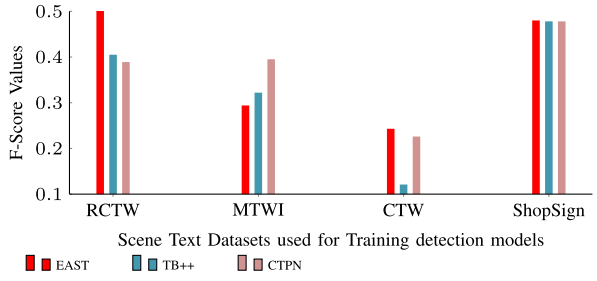

Fig. 8. F-scores of street view text detection algorithms EAST, TB++, CTPN trained on different datasets and tested on ShopSign (Multi-Oriented).

Train1 of ShopSign, the three scene text detection algorithms obtained very close detection results on the multi-oriented images of Test1. Therefore, for DL based scene text detection algorithms, their performance depends both on the internal mechanism of the algorithms and the specific datasets used in training. In other words, the performance of DL based scene text detections algorithms are both algorithm-dependent and data-dependent.

2) Extensive Scene Text Detection Experiments: In order to eliminate the influence of dataset split (*Split-1*) on the evaluation of scene text detection algorithms, we use *Split-2* to re-run the experiments with the same scene text detection algorithms EAST, TB++ and CTPN. Meanwhile, we include three latest algorithms which achieved very good text detection performance on the English datasets, which are TextSnake[6] [68], DeRPN[7] [69] and TextField[8] [70]. Here, we use four well-known datasets that contain Chinese, which are MSRA-TD500, RCTW, MTWI, and MLT, to test the detection performance with/without the ShopSign dataset. For RCTW, MTWI, and MLT, since their official test data are not released, we only use their original training data, which is randomly split into train set and tests in our experiments. The dataset splits for RCTW and MTWI have been mentioned above; for MLT, we use 7,200 images for training and 1,800 images for testing (the numbers of samples for each language are the same).

In Table IV, we report the results of these six algorithms. The threshold for IoU is 0.5. First of all, we notice that adding the ShopSign dataset does not always help improve scene text detection performance, depending on the specific algorithms. For EAST and TextSnake, the join of ShopSign can enhance the text detection performance; yet, for TextBoxes++, TextSnake, DeRPN and CTPN, the influence is negative. We also find that, these algorithms have different performance on different datasets, as can be seen from Figure 9(standalone datasets, without adding ShopSign). For instance, the best performing algorithms on MSRA-TD500[9] are TextBoxes++ and TextField, while on MLT, the best performers are EAST and TextField. Overall, EAST, TextBoxes++ and TextSnake are the top-3 algorithms on scene text detection. TextField and TextBoxes++ always achieve the best precision, while the recall of DeRPN is commonly the highest among all the algorithms.

Comparing the results in Table IV and Table III, we see that the overall text detection performance of EAST and TextBoxes++ do not vary much.[10] This also verifies that *Split-1* on ShopSign can reflect the challenges of ShopSign.

3) Text Detection Performance on the Five Categories of Hard Images: To show the detection difficulty on the five

[6]Source codes: https://github.com/princewang1994/TextSnake.pytorch.

[7]Source codes: https://github.com/HCIILAB/DeRPN.

[8]Source codes: https://github.com/YukangWang/TextField.

[9]The detection performance on MSRA-TD500 by EAST is 0.761 in the original paper, while our experimental result was 0.549 in this article. In the original EAST paper, EAST included another 400 images from HUST-TR400 dataset as the training data, thus the final training data is much larger than the original one. This partially explains the above difference in accuracy. Another reason lies in the parameter fine-tuning process.

[10]For CTPN, since *Split-1* contains more horizontal text instances than *Split-2* (because of the 2,516 pairs of images), the results on the splits are different.

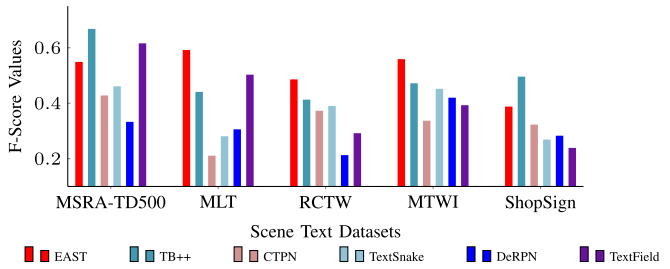

Fig. 9. F-scores of scene text detection algorithms EAST, TB++, CTPN, TextSnake, DeRPN, and TextField on MSRA-TD500, MLT, RCTW, MTWI, and ShopSign.

"hard" categories of ShopSign, which are mirror, wooden, deformed, exposed and obscured, we report the specific text detection results on each category, using *Split-3* (Train3 and Test3) of ShopSign. But the performance difference from what we observe in Table III is not significant. This is because the results are image-level, they include all the text-lines for images that contain hard text instances (such as the ones on the mirrors).

To further check the performance of scene detection algorithms on the specific "hard" examples of ShopSign images, in Table V, we only pick the corresponding "hard" text instances from each "hard" image, and separately calculate their recall. That is, for each image that belongs to the "hard" category, we only measure the recall results of the specific hard text instances of the image. It must be noted that, with this special evaluation strategy, it is only possible to calculate the recall score, but not the precision score, because for precision it is impossible to distinguish the corresponding prediction results (by the scene text detectors) for the "hard" text instances in each "hard" image. Thus, the denominator of the precision formula can not be determined. But for the recall formula, the denominator is the number of ground-truth hard text instances, while the numerator is the number of hard text instances that can be detected. Hence, we only report the recall results of the scene text detection algorithms on the hard text instances of ShopSign, which is presented in Table V.

It is clear from Table V that the recall results for the hard text instances in ShopSign are much lower than those reported in Table III. This shows that, the five specific "hard" examples of ShopSign are more challenging. Moreover, we see that TextBoxes++ is less influenced by the hard examples than EAST and CTPN. In particular, both EAST and CTPN perform poorly on the mirror and obscured images, with a recall under 41.2%. We also notice that, out of the five categories of "hard" images, the wooden category of ShopSign is the least difficult to these scene text detection algorithms.

C. Scene Text Recognition Performance

Chinese scene text recognition is a largely neglected topic, mainly due to the high complexity of the problem and the lack of large-scale datasets in the earlier years. In the following, we will report the performance of existing scene text recognition algorithms on ShopSign and the other datasets.

1) Edit Distance Accuracy: Now we present scene text recognition results on ShopSign, using *Split-1*. For reference, we also include our results on three other Chinese datasets, which are RCTW, MTWI, and CTW. As mentioned above, an interesting and unique characteristic of ShopSign is that it equally splits the test images into two sets; one containing images with horizontal texts and the other with multi-oriented texts. This enables us to separately study the recognition performance of existing algorithms on horizontal and multi-oriented (perspective) texts. The text recognition methods used for this study are ASTER [15], CRNN [13], SlidingCNN [14] and SliceCNN. We note that our main aim here is not to compare the performance of text recognition algorithms but to report their accuracy results on ShopSign.

As can be seen from Table VI, ASTER yields an accuracy of 63% and 54% on the horizontal and multi-oriented test sets of ShopSign, respectively. This difference should be expected, since multi-oriented text-lines may contain perspective text instances that are more difficult to be recognized. In contrast, the other methods fall significantly behind ASTER, as can be observed from Figure 10, with an accuracy of less than 25% on ShopSign. ASTER obtains an accuracy of 73% on MTWI, which is the highest accuracy for all the algorithms. This is due to the fact that most text-line instances in MTWI are horizontal and many of these crops are born-digital and with relatively clean backgrounds, thus they are less difficult to be recognized.

TABLE V

RECALL (SPECIFIC TEXT-LINE LEVEL) OF EAST, TB++, AND CTPN ON THE FIVE HARD CATEGORIES OF MIRROR, WOODEN, DEFORMED (DEF.), EXPOSED (EXP.), AND OBSCURED (OBS.)

Methods	Datasets	Mirror	Wood	Def.	Exp.	Obs.
EAST	CTW	0.096	0.152	0.201	0.239	0.112
	+ShopSign	0.376	0.488	0.462	0.543	0.341
	MTWI	0.196	0.264	0.316	0.351	0.155
	+ShopSign	0.388	0.492	0.496	0.564	0.343
	RCTW	0.296	0.389	0.444	0.452	0.278
	+ShopSign	0.380	0.492	0.479	0.585	0.359
TB++	CTW	0.244	0.373	0.333	0.282	0.302
	+ShopSign	0.488	0.635	0.427	0.521	0.473
	MTWI	0.356	0.450	0.359	0.372	0.320
	+ShopSign	0.524	0.645	0.466	0.548	0.486
	RCTW	0.436	0.535	0.380	0.468	0.402
	+ShopSign	0.500	0.637	0.432	0.569	0.482
CTPN	CTW	0.096	0.177	0.248	0.133	0.124
	+ShopSign	0.332	0.373	0.350	0.362	0.342
	MTWI	0.340	0.401	0.346	0.356	0.328
	+ShopSign	0.412	0.455	0.368	0.415	0.372
	RCTW	0.368	0.420	0.359	0.426	0.366
	+ShopSign	0.372	0.420	0.372	0.367	0.407

TABLE VI

TEXT RECOGNITION ACCURACY ON SHOPSIGN, RCTW, MTWI AND CTW, USING ASTER, CRNN, SLIDINGCNN, AND SLICECNN. HOR AND MUL DENOTE THE ACCURACY RESULTS ON IMAGES WITH HORIZONTAL AND MULTI-ORIENTED TEXT-LINES, RESPECTIVELY

Algorithms	RCTW	MTWI	CTW	ShopSign	
				Hor	Mul
ASTER	**0.44**	**0.73**	0.27	**0.63**	**0.54**
CRNN	0.25	0.35	**0.28**	0.25	0.19
SlidingCNN	0.19	0.37	0.09	0.22	0.14
SliceCNN	0.16	0.37	0.10	0.27	0.18

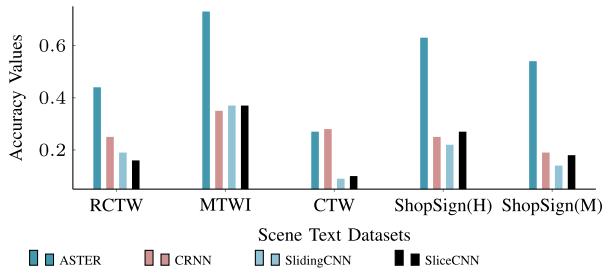

Fig. 10. Edit distance accuracy results of street view text recognition algorithms ASTER, CRNN, SlidingCNN, SliceCNN on RCTW, MTWI, CTW, and ShopSign.

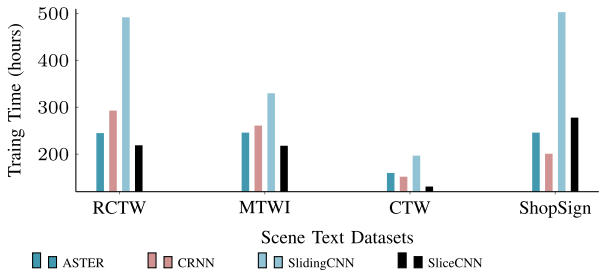

Fig. 11. Training time of street view text recognition algorithms ASTER, CRNN, SlidingCNN, SliceCNN on RCTW, MTWI, CTW, and ShopSign.

TABLE VII
TEXT RECOGNITION ACCURACY ON THE HARD CATEGORIES OF SHOPSIGN USING THE TEXT RECOGNITION MODELS TRAINED ON MTWI AND RCTW WITH ASTER

Model	Mir.	Wood	Def.	Exp.	Obs.	ShopSign	
						Hor	Mul
MTWI	0.233	0.361	0.281	0.177	0.218	0.304	0.439
RCTW	0.177	0.288	0.234	0.172	0.094	0.247	0.344

CRNN is found to be the second in accuracy behind ASTER, but it yields 2 times lower rates than ASTER on RCTW, MTWI, and ShopSign, but obtains very close results as ASTER on CTW. SlidingCNN comes last in performance, trailing CRNN in all the datasets except MTWI. Its accuracy rates on CTW and ShopSign are only $9\% \sim 22\%$, respectively, which are low for practical use. We also see that the accuracy results of SliceCNN and SlidingCNN are similar. On CTW and ShopSign, SliceCNN obtains better recognition accuracy than SlidingCNN. We also see that, both algorithms outperform CRNN on the MTWI dataset.

In Table VII, we show the recognition results on the five hard categories. For more intuitive comparisons between the accuracies of the "ordinary" text instances (i.e., the text instances in Test1 of *Split-1*) and the special "hard" text instances (i.e., the text instances in Test3 of *Split-3*), we utilize the "third-party" models trained respectively on the MTWI and RCTW datasets using the ASTER algorithm, and compare their performance on the "ordinary" text instances in ShopSign and the "hard" text instances of the five special categories. We see that, the recognition accuracy of DL based scene text recognition algorithms are typically lower on the hard categories. In particular, on the *exposed*, *obscured* and *mirror* categories, the accuracy drops significantly. There is only an exception with the *wooden* category, where the recognition accuracy was not influenced. Overall, it is more challenging to recognize texts in the hard categories of ShopSign.

2) Efficiency: In Table VIII, we report the efficiency of different scene text recognition algorithms, and the sizes of the corresponding models. We see that ASTER consumes the largest amount of space, while SliceCNN and SlidingCNN occupy the smallest space. In terms of model training time, as can be seen from Figure 11, the proposed SliceCNN method is the fastest in training (with an exception on ShopSign). Regarding model prediction speed, it is clear that ASTER is the fastest on all the datasets, followed by SliceCNN.

Overall, ASTER is generally the top performer in both scene text recognition accuracy and prediction speed; but it also requires the largest amount of model space. However, given the low accuracy results on RCTW, CTW and ShopSign, the recognition performance of scene text recognition algorithms still have a very large room for further improvements on the Chinese scene text datasets, to fit the requirements of accurate urban scene understanding in ITS.

We note that, not all the scene images contain texts. For urban scene understanding in ITS applications, for non-text images researchers can still resort to conventional approaches for scene understanding (such as image captioning); but for images with texts, researchers should combine conventional approaches with explicit text information extracted using Photo OCR approaches to better understand the scenes in ITS, which has a very high requirement in the accuracy of scene understanding.

V. CROSS-LANGUAGE STREET VIEW TEXT DETECTION

For street view text recognition, it is clear that training data with the same language/script is needed to train the corresponding text recognition model. However, for street view text detection, it only needs to localize the textual areas in the scene images, but without recognizing the texts. It is thus natural to raise the following question: is street view text detection language/script dependent?

Cross-language text detection is a quite meaningful research topic as language factor has not been well examined in a systematic manner in the scene text detection field. Researchers are aware that languages matter but not sure how much they matter. To investigate this issue, we train text detection models on the English natural scene images and test their performance on the Chinese scene images, and vice versa.

A. Train Models on English Scene Images and Test Them on the Chinese Ones

Let E be the set of models trained from English natural scene images, and C be the set of models trained from the Chinese ones. For E, we directly use the pre-trained models that are publicly available [11], [12]. But for C, we train the models from scratch, since there are no such pre-trained models for Chinese images. We note that, for fair comparisons,

TABLE VIII

MODEL SIZE, TRAINING TIME AND TESTING SPEED OF THE TEXT RECOGNITION ALGORITHMS OVER DIFFERENT CHINESE SCENE TEXT DATASETS

Algorithms	Model Size (MB)				Training Time (Hours)				Testing / Prediction Speed (Sec/Frame)			
	RCTW	MTWI	CTW	ShopSign	RCTW	MTWI	CTW	ShopSign	RCTW	MTWI	CTW	ShopSign
ASTER	481	518	492	506	245	246	160	246	**0.068**	**0.072**	**0.069**	**0.068**
CRNN	77	77	77	77	293	261	152	**201**	0.608	0.633	0.345	0.282
SlidingCNN	59	59	59	59	492	330	197	503	0.752	0.830	0.762	0.838
SliceCNN	**59**	**59**	**59**	**59**	**219**	**218**	**131**	278	0.441	0.542	0.460	0.515

TABLE IX

TEXT DETECTION PERFORMANCE OF THE PRE-TRAINED ENGLISH SCENE TEXT DETECTION MODELS ON THE CHINESE SCENE TEXT DATASETS USING EAST AND TEXTBOXES++ (TB++). R, P, AND H ARE SHORT NAMES FOR RECALL, PRECISION AND F-SCORE, RESPECTIVELY

Train	Test	Algo.	Text Detection Performance		
			R	P	H
Model E_1					
ICDAR 13+15	RCTW	EAST	0.243	0.265	0.253
ICDAR 13+15	MTWI	EAST	0.287	0.444	0.348
ICDAR 13+15	CTW	EAST	0.158	0.161	0.159
ICDAR 13+15	ShopSign	EAST	0.129	0.146	0.137
Model E_2					
SynthText	RCTW	TB++	0.176	0.295	0.221
SynthText	MTWI	TB++	0.238	0.510	0.325
SynthText	CTW	TB++	0.030	0.178	0.051
SynthText	ShopSign	TB++	0.132	0.310	0.185

we are supposed to train E and C on the same number of samples, but we opt to reuse the pre-trained models to roughly demonstrate the influence of language on scene text detection.

Specifically, we first use the model E_1 pre-trained on both ICDAR 2013 and ICDAR 2015 using EAST [11], and check the performance on the four Chinese datasets RCTW, MTWI, CTW, and ShopSign, respectively. We would like to note that, until now, there is little work that has studied the detection performance of the English text detection models on the Chinese datasets. In [11], the reported recall and precision rates of the model E_1 on ICDAR 2015 are 78.3% and 83.3%, respectively. However, as shown in Table IX, E_1 yields very low detection performance on the four Chinese datasets. On the test set of ShopSign, the recall and precision rates are only 12.9% and 14.6%, respectively, which is significantly lower than the detection performance yielded on the same English datasets [11]. Similar observations can be found on the other Chinese datasets, their recall results range from 15.8% to 28.7%. These results also reflect the low generalization capability of DL based scene detection algorithms/models.

Next, we adopt the model E_2 pre-trained on SynthText that contains around 800,000 English synthetic images [35] using TextBoxes++ [12], to test its detection performance on the four Chinese datasets. In [12], the reported recall and precision of TextBoxes++ model E_2 on ICDAR 2013 are 84% and 91%, respectively. However, as depicted in Table IX, the recall of this model on ShopSign is only 13.2%, which is also very low for practical use. Similar observations can be found on MTWI, CTW and RCTW.

In short, regardless of the datasets, models trained on English scene images are unable to effectively localize Chinese scene texts.

TABLE X

TEXT DETECTION PERFORMANCE OF CHINESE MODELS ON THE ENGLISH SCENE TEXT DATASETS USING EAST AND TEXTBOXES++ (TB++)

Data	ICDAR 2013			ICDAR 2015		
	R	P	H	R	P	H
EAST (the corresponding models are named C_1 to C_4)						
RCTW	0.310	0.535	0.392	0.299	0.511	0.378
MTWI	0.497	0.645	0.561	0.398	0.515	0.449
CTW	0.021	0.338	0.040	0.019	0.381	0.040
ShopSign	0.284	0.473	0.355	0.264	0.467	0.338
TB++ (the corresponding models are named C_5 and C_6)						
RCTW	0.392	0.592	0.471	0.249	0.370	0.298
ShopSign	0.428	0.526	0.472	0.288	0.304	0.296

B. Train Models on Chinese Scene Images and Test Them on the English Ones

Reversely, we check the performance of C (the set of models trained from the Chinese scene text datasets) on the English images of ICDAR 2013 and ICDAR 2015. The recall and precision results are presented in Table X. We see that, the Chinese scene text detection models trained on RCTW and ShopSign (C_1 and C_4) using EAST, only obtain a recall between 26% and 31% on both ICDAR datasets. The text detection model trained on MTWI (C_2) using EAST yields the best overall detection performance, which is possibly due to the fact that MTWI contains a large amount of English scripts, thus models trained on this dataset are more capable in English text detection. TextBoxes++ achieves better recall results on the ICDAR 2013 dataset than EAST when using C_5 and C_6, but this does not hold on ICDAR 2015. Overall, models trained on Chinese scene text datasets can not effectively localize texts in the English natural scene images.

C. Training and Testing on Different Chinese Scene Text Datasets

Moreover, in Table XI, we report the detection performance of a model trained on one Chinese scene text dataset and tested on another different Chinese scene text dataset. By comparing Tables IX and XI, we see that by training models from the Chinese datasets, we are able to achieve significantly better performance on the Chinese scene images than the models trained on the English datasets. It is also interesting to notice that EAST always outperforms TextBoxes++ in terms of recall on all the Chinese datasets.

Through our observations from Tables IX, X and XI, we deduce that script language does have a significant influence on the performance of scene text detection algorithms. For Chinese scene text detection, it is therefore essential to

TABLE XI
TRAIN AND TEST TEXT DETECTION MODELS ON CHINESE SCENE TEXT DATASETS USING EAST AND TEXTBOXES++ (TB++). FOR SHOPSIGN, DUE TO SPACE LIMIT WE ONLY REPORT THE DETECTION PERFORMANCE ON THE MULTI-ORIENTED IMAGES

Train	Test	Algo.	Detection Performance		
			R	P	H
ShopSign	ShopSign	EAST	0.579	0.410	0.480
RCTW	RCTW	EAST	0.491	0.508	0.499
MTWI	MTWI	EAST	0.428	0.787	0.554
CTW	CTW	EAST	0.547	0.581	0.563
ShopSign	RCTW	EAST	0.488	0.482	0.485
		TB++	0.430	0.471	0.449
RCTW	ShopSign	EAST	0.439	0.355	0.392
		TB++	0.395	0.555	0.461
ShopSign	MTWI	EAST	0.269	0.591	0.370
		TB++	0.327	0.552	0.411
RCTW	MTWI	EAST	0.340	0.260	0.295
		TB++	0.302	0.631	0.409

use Chinese natural scene datasets to train the street view text detection models. More advanced street view text detection algorithms are also needed.

VI. DISCUSSIONS

Street view text recognition techniques have three major functions to intelligent transportation systems and autonomous driving: (a) urban scene understanding during driving, because texts in the images provide explicit information for image and environment understanding; (b) automatic POIs collection for digital and HD mapping; (c) precise positioning, since the vast amount of extracted shop sign entities can be used as anchor points in positioning computation. However, little effort has been devoted to this direction. This motivates us to investigate street view text recognition techniques and utilize them in ITS.

Concerning the experimental results, we have the following summary of findings:

(1). Chinese street view texts recognition is significantly more difficult than the English ones. Existing methods yield an accuracy of more than 90% on many English benchmark datasets, but their overall accuracy on ShopSign is about 60%. Therefore, there is a great room for further improvements to meet real-world ITS applications, which have very high requirement on the accuracy and robustness of the deployed algorithms.

(2). Language has an important influence on scene text detection, although it does not need to identify the texts in the images but only locates the text areas. OCR researchers are aware that languages matter but are not sure how much they matter. From our experimental results in Section V, using the same EAST model trained on ICDAR datasets that obtained 78.3% and 83.3% on ICDAR 2015 in terms of recall and precision, it only yields 12.9% in recall and 14.6% in precision on ShopSign, which is significantly lower than on the English datasets. Likewise, the EAST text detection model trained on ShopSign that achieved a recall of 57.9%, only obtained a recall of less than 28.4% on the ICDAR datasets. Therefore, cross-language and multi-lingual scene text detection and recognition should be a useful research direction with lots of challenges and opportunities.

(3). Scene text detection algorithms still have performance bottleneck on Chinese street view scene texts. State-of-the-art methods only obtain an F-score of less than 50% on ShopSign, which can not be used in critical applications such as ITS and autonomous driving.

(4). Since both the detection and recognition performance on ShopSign is not very satisfactory, people should either make efforts to build larger benchmark datasets, or develop new methodologies for street view text recognition, other than continue using the current sequence-to-sequence learning approaches. Character-level scene text detection and recognition techniques have good potential in this direction and deserve more devotion.

VII. CONCLUSION

Many researchers are aware that, without texts, it is often ambiguous to understand the meanings in the images, whereas texts extracted by OCR algorithms from the scene images provide explicit evidence for accurate scene understanding. However, little attention has been paid to the utilization of street view text recognition algorithms for scene understanding in ITS and safe autonomous driving in city scenarios.

To address this problem, we introduce ShopSign, which is a large-scale dataset specialized in Chinese signboards in street views. We carry out an extensive empirical study on the performance of existing scene text detection and recognition algorithms on ShopSign and three other Chinese scene text datasets. We reveal that language/script has a strong influence on scene text detection. Moreover, we empirically compare the accuracy of existing scene text recognition algorithms. ASTER is found to be the state-of-the-art method in Chinese scene text recognition, but there is still a large room for further improvements. Overall, ShopSign is a new benchmark dataset with lots of new challenges and research opportunities to be exploited in street view scene text reading. The extensive experimental studies and results from the work should also have important reference value to the community.

In future work, we will investigate how to combine scene text recognition and image captioning to enhance scene image understanding. Besides ITS and autonomous driving, in future applications such as 3D city reconstruction [71] from multiple heterogeneous data resources including satellite images, low-altitude 3D aerial images and street-view scene images, street view text recognition and generation techniques should be an important functioning module.

ACKNOWLEDGMENT

This research is very difficult, it lasted more than three years and involved about 50 personnel (including researchers, graduate students, and undergraduates). It took us more than two years to build this large-scale street view ShopSign dataset, and over one year to design the scene text recognition algorithms and run the extensive experiments and comparisons. First of all, the authors are very grateful to all the research assistants who contributed to the collection/annotation of the ShopSign

dataset. Besides, the first author himself has also taken a very large amount of representative photos with texts, during his stays/missions at Shanghai, Inner Mongolia, Beijing, Xiamen, Zhengzhou and KaiFeng. Second, the authors will never forget the great help of Mr. Yuefeng Tao, Mr. Qin Niu and Mr. Menglei Jiao in some of the experiments of this work. Third, the authors would like to thank Prof. Xu-Cheng Yin (USTB), Dr. Chun Yang (USTB), Mr. Chang Liu (USTB), Dr. Ke Chen (SCUT), Dr. Wei Jiang (NCWU), and Dr. George Almpanidis (HENU) for their discussions and help. Finally, we thank all the anonymous reviewers for their insightful comments and suggestions.

REFERENCES

[1] H. G. Seif and X. Hu, "Autonomous driving in the iCity—HD maps as a key challenge of the automotive industry," *Engineering*, vol. 2, no. 2, pp. 159–162, Jun. 2016.

[2] K. Jiang, D. Yang, C. Liu, T. Zhang, and Z. Xiao, "A flexible multi-layer map model designed for lane-level route planning in autonomous vehicles," *Engineering*, vol. 5, no. 2, pp. 305–318, Apr. 2019.

[3] D. Feng *et al.*, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transport. Syst.*, pp. 1–20, 2020.

[4] X. Bai, M. Yang, P. Lyu, Y. Xu, and J. Luo, "Integrating scene text and visual appearance for fine-grained image classification," *IEEE Access*, vol. 6, pp. 66322–66335, 2018.

[5] H. Yin, Y. Wang, X. Ding, L. Tang, S. Huang, and R. Xiong, "3D LiDAR-based global localization using Siamese neural network," *IEEE Trans. Intell. Transport. Syst.*, vol. 21, no. 4, pp. 1380–1392, Apr. 2020.

[6] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3D object detection methods for autonomous driving applications," *IEEE Trans. Intell. Transport. Syst.*, vol. 20, no. 10, pp. 3782–3795, Oct. 2019.

[7] D. Tabernik and D. Skocaj, "Deep learning for large-scale traffic-sign detection and recognition," *IEEE Trans. Intell. Transport. Syst.*, vol. 21, no. 4, pp. 1427–1440, Apr. 2020.

[8] Y. Jin, X. Guo, Y. Li, J. Xing, and H. Tian, "Towards stabilizing facial landmark detection and tracking via hierarchical filtering: A new method," *J. Franklin Inst.*, vol. 357, no. 5, pp. 3019–3037, Mar. 2020.

[9] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. ECCV*, 2016, pp. 56–72.

[10] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 745–753.

[11] X. Zhou *et al.*, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2642–2651.

[12] 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.

[13] 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, Nov. 2017.

[14] F. Yin, Y.-C. Wu, X.-Y. Zhang, and C.-L. Liu, "Scene text recognition with sliding convolutional character models," 2017, *arXiv:1709.01727*. [Online]. Available: http://arxiv.org/abs/1709.01727

[15] 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.

[16] C. Zhang *et al.*, "ShopSign: A diverse scene text dataset of Chinese shop signs in street views," 2019, *arXiv:1903.10412*. [Online]. Available: http://arxiv.org/abs/1903.10412

[17] 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. (ICML)*, 2006, pp. 369–376.

[18] 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.

[19] 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.

[20] Y. Zhu, C. Yao, and X. Bai, "Scene text detection and recognition: Recent advances and future trends," *Frontiers Comput. Sci.*, vol. 10, no. 1, pp. 19–36, Feb. 2016.

[21] X.-C. Yin, Z.-Y. Zuo, S. Tian, and C.-L. Liu, "Text detection, tracking and recognition in video: A comprehensive survey," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2752–2773, Jun. 2016.

[22] 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.

[23] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," 2018, *arXiv:1811.04256*. [Online]. Available: http://arxiv.org/abs/1811.04256

[24] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions," in *Proc. 7th Int. Conf. Document Anal. Recognit. (ICDAR)*, 2003, pp. 682–687.

[25] D. Karatzas *et al.*, "ICDAR 2013 robust reading competition," in *Proc. 12th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2013, pp. 1484–1493.

[26] D. Karatzas *et al.*, "ICDAR 2015 competition on robust reading," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1156–1160.

[27] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1083–1090.

[28] N. Nayef *et al.*, "ICDAR2017 robust reading challenge on multi-lingual scene text detection and script Identification–RRC-MLT," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, Nov. 2017, pp. 1454–1459.

[29] S. Tian *et al.*, "Multilingual scene character recognition with co-occurrence of histogram of oriented gradients," *Pattern Recognit.*, vol. 51, pp. 125–134, Mar. 2016.

[30] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1457–1464.

[31] A. Mishra, K. Alahari, and C. Jawahar, "Scene text recognition using higher order language priors," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2012, pp. 1–11.

[32] B. Shi *et al.*, "ICDAR2017 competition on reading Chinese text in the wild (RCTW-17)," in *Proc. ICDAR*, 2017, pp. 1429–1434.

[33] M. He *et al.*, "ICPR2018 contest on robust reading for multi-type Web images," in *Proc. ICPR*, 2018, pp. 7–12.

[34] T.-L. Yuan, Z. Zhu, K. Xu, C.-J. Li, T.-J. Mu, and S.-M. Hu, "A large Chinese text dataset in the wild," *J. Comput. Sci. Technol.*, vol. 34, no. 3, pp. 509–521, May 2019.

[35] 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.

[36] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[37] 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.

[38] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.

[39] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[40] 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*, 2017, pp. 4161–4167.

[41] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *Proc. ECCV*, 2016, pp. 21–37.

[42] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[43] 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.

[44] 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. 9365–9374.

[45] S.-X. Zhang *et al.*, "Deep relational reasoning graph network for arbitrary shape text detection," in *Proc. IEEE CVPR*, Jun. 2020, pp. 9699–9708.
[46] D. Yu *et al.*, "Towards accurate scene text recognition with semantic reasoning networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12113–12122.
[47] 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. 9809–9818.
[48] 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. (BMVC)*, 2016, pp. 1–7.
[49] 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. AAAI*, 2018, pp. 7194–7201.
[50] Z. Qiao, Y. Zhou, D. Yang, Y. Zhou, and W. Wang, "SEED: Semantics enhanced encoder-decoder framework for scene text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13528–13537.
[51] 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. 11962–11972.
[52] 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. (IJCAI)*, Aug. 2017, pp. 3280–3286.
[53] 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. (CVPR)*, Jun. 2018, pp. 5571–5579.
[54] 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*, 2019, pp. 8610–8617.
[55] F. Zhan, C. Xue, and S. Lu, "GA-DAN: Geometry-aware domain adaptation network for scene text detection and recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9105–9115.
[56] 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. 2059–2068.
[57] Z. Wan, M. He, H. Chen, X. Bai, and C. Yao, "Textscanner: Reading characters in order for robust scene text recognition," in *Proc. AAAI*, 2020, pp. 12120–12127.
[58] L. Xing, Z. Tian, W. Huang, and M. Scott, "Convolutional character networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9126–9136.
[59] M. Liao *et al.*, "Scene text recognition from two-dimensional perspective," in *Proc. AAAI*, 2019, pp. 8714–8721.
[60] J. Baek *et al.*, "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.
[61] X. Wu, S. Dai, Y. Guo, and H. Fujita, "A machine learning attack against variable-length Chinese character CAPTCHAs," *Int. J. Speech Technol.*, vol. 49, no. 4, pp. 1548–1565, Apr. 2019.
[62] X. Xu, J. Chen, J. Xiao, L. Gao, F. Shen, and H. T. Shen, "What machines see is not what they get: Fooling scene text recognition models with adversarial text images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12304–12314.
[63] L. Wu *et al.*, "Editing text in the wild," in *Proc. ACM Multimedia*, 2019, pp. 1500–1508.
[64] C. Luo, Y. Zhu, L. Jin, and Y. Wang, "Learn to augment: Joint data augmentation and network optimization for text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13746–13755.
[65] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7699–7707.
[66] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[67] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 3354–3361.
[68] 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. ECCV*, 2018, pp. 20–36.
[69] L. Xie, Y. Liu, L. Jin, and Z. Xie, "DeRPN: Taking a further step toward more general object detection," in *Proc. AAAI*, 2019, pp. 9046–9053.
[70] Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang, and X. Bai, "TextField: Learning a deep direction field for irregular scene text detection," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5566–5579, Nov. 2019.
[71] L. Duan and F. Lafarge, "Towards large-scale city reconstruction from satellites," in *Proc. ECCV*, 2016, pp. 89–104.

Chongsheng Zhang (Member, IEEE) received the Ph.D. degree from INRIA, France. He was a Visiting Scholar with UCLA, Los Angeles, USA, under Prof. Carlo Zaniolo, from 2010 to 2011. He worked as an ERCIM Marie Curie Fellow with NTNU, Trondheim, Norway, from 2012 to 2013. He is currently a Full Professor with Henan University, Kaifeng, China, where he also leads the Data Science and Artificial Intelligence (DSAI) Research Team. He has published more than 30 articles in peer-reviewed journals and conferences. He holds six Chinese innovation patents and filed another four Chinese patent applications. He has authored five books in data science and artificial intelligence, including big data analytics, artificial intelligence: face recognition and retrieval, and deep learning: theories and applications. His research interests include machine learning, deep learning, and OCR. He is an Associate Editor of IEEE ACCESS for the period of 2019 to 2022 and a PC Member for ECML-PKDD 2018, CIKM 2020, and CBMS from 2014 to 2020. He is also a Reviewer for many well-known international journals such as *Knowledge-Based Systems*, *Information Sciences*, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and the IEEE TRANSACTIONS ON RELIABILITY.

Weiping Ding (Senior Member, IEEE) received the Ph.D. degree in computation application from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2013. He was a Visiting Scholar with the University of Lethbridge (UL), Alberta, Canada, in 2011. From 2014 to 2015, he was a Post-Doctoral Researcher with the Brain Research Center, National Chiao Tung University (NCTU), Hsinchu, Taiwan. In 2016, he was a Visiting Scholar with the National University of Singapore (NUS), Singapore. From 2017 to 2018, he was a Visiting Professor with the University of Technology Sydney (UTS), Ultimo, NSW, Australia. He has published more than 80 research peer-reviewed journal articles and conference papers, including the IEEE TRANSACTIONS ON FUZZY SYSTEMS (TFS), the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), the IEEE TRANSACTIONS ON CYBERNETICS (TCYB), the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS (TSMCS), the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING (TBME), the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS (TII), the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE (TETCI), and the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (TITS). His main research interests include data mining, granular computing, evolutionary computing, machine learning, and big data analytics. He is a member of the IEEE CIS, ACM, and CCAI, and a Senior CCF Member. He is a member of Technical Committee on Soft Computing of the IEEE SMCS, Granular Computing of the IEEE SMCS, and Data Mining and Big Data Analytics of the IEEE CIS. He is the Chair of the IEEE CIS Task Force on Granular Data Mining for Big Data. He currently serves on the Editorial Advisory Board of *Knowledge-Based Systems* and Editorial Board of *Information Fusion* and *Applied Soft Computing*. He serves/served as an Associate Editor for the IEEE TRANSACTIONS ON FUZZY SYSTEMS, *Information Sciences*, *Swarm and Evolutionary Computation*, IEEE ACCESS, and the *Journal of Intelligent and Fuzzy Systems*, and the Co-Editor-in-Chief for the *Journal of Artificial Intelligence and System*. He is the Leading Guest Editor of Special Issues in several prestigious journals, including the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, the IEEE TRANSACTIONS ON FUZZY SYSTEMS, *Information Fusion*, and *Information Sciences*. He has delivered more than 15 keynote speeches at international conferences and has co-chaired several international conferences and workshops in the areas of data mining, fuzzy decision-making, and knowledge engineering.

Guowen Peng received the B.Sc. and master's degrees in computer science from Henan University in 2016 and 2019, respectively. Her research interests include scene text recognition and machine learning.

Feifei Fu received the B.Sc. degree in software engineering from Henan University in 2017, where she is currently pursuing the master's degree. Her research interests include scene text recognition and machine learning.

Wei Wang received the Ph.D. degree in control science and engineering from Xi'an Jiaotong University in 2006. He was a Post-Doctoral Researcher with the University of Trento, Italy, from 2005 to 2006. He was a Post-Doctoral Researcher with TELECOM Bretagne and INRIA, France, from 2007 to 2008. He was a European ERCIM Fellow with the Norwegian University of Science and Technology (NTNU), Norway, and the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, from 2009 to 2011. He is currently a Full Professor and the Head of the Department of Information Security, Beijing Jiaotong University, China. He has authored or coauthored over 90 peer-reviewed articles in various journals and international conferences. His main research interests include mobile, computer, and network security. He is an Editorial Board Member of *Computers and Security* and the Young AE of Frontiers of *Computer Science*.