Challenges of Deep Learning-based Text Detection in the Wild

Zobeir Raisi — Vision and Image Processing Lab, University of Waterloo, ON, N2L 3G1, Canada
Mohamed A. Naiel — Vision and Image Processing Lab, University of Waterloo, ON, N2L 3G1, Canada
Paul Fieguth — Vision and Image Processing Lab, University of Waterloo, ON, N2L 3G1, Canada
Steven Wardell — ATS Automation Tooling Systems Inc., Cambridge, ON, N3H 4R7, Canada
John Zelek — Vision and Image Processing Lab, University of Waterloo, ON, N2L 3G1, Canada
Email: {zraisi, mohamed.naiel, pfieguth, jzelek}@uwaterloo.ca, swardell@atsautomation.com

Abstract

The reported accuracy of recent state-of-the-art text detection methods, mostly deep learning approaches, is in the order of 80% to 90% on standard benchmark datasets. These methods have relaxed some of the restrictions of structured text and environment (i.e., "in the wild") which are usually required for classical OCR to properly function. Even with this relaxation, there are still circumstances where these state-of-the-art methods fail. Several remaining challenges in wild images, like in-plane-rotation, illumination reflection, partial occlusion, complex font styles, and perspective distortion, cause exciting methods to perform poorly. In order to evaluate current approaches in a formal way, we standardize the datasets and metrics for comparison which had made comparison between these methods difficult in the past. We use three benchmark datasets for our evaluations: ICDAR13, ICDAR15, and COCO-Text V2.0. The objective of the paper is to quantify the current shortcomings and to identify the challenges for future text detection research.

1 Introduction

Detecting and recognizing text in the wild images are challenging problems in the field of computer vision [1, 2]. "In the wild" refers to problems where the structured environment and text are of wide variations. Examples include street signs, store signs, advertisements, or text identifying sport players, to name a few. Reading text from scene images can be carried-out using two fundamental tasks:

- Text detection that localizes text in the image, and
- Text recognition that converts localized text or a cropped word image into a text string.

They face common challenging problems that can be categorized as:

- **Text diversity:** images that contain text with different colors, fonts, orientations and languages.
- **Scene complexity:** images that include scene elements of similar appearance to text, such as signs, bricks and symbols.
- **Distortion factors:** text images distorted due to the effect of motion blurriness, images of low resolution, surface geometry, perspective distortion and partial occlusion [1, 3, 4].

This paper focuses on the text detection task, which is more challenging than text recognition due to the large variance of text shape and complicated backgrounds. The methods before the deep learning era, typically identify a character or text component candidates using connected component-based approaches or sliding window-based methods, which used hand-craft features like MSER [5] or SWT [6] as basic components. However, the detection performance of these classical machine learning-based methods is still far from satisfactory.

Recently, deep learning-based methods have been shown to outperform in detecting challenging text in scene images. These methods usually adopt general object detection frameworks such as SSD [7], YOLO [8], Faster R-CNN [9], or segmentation frameworks like FCN [10] and Mask R-CNN [11]. Most deep learning-based text detectors that detect text at the word level have difficulties in finding curved, extremely long, or highly deformed words by using a single bounding box [12].

This paper aims to highlight on the preceding challenges by reviewing recent advances in deep learning applied to scene text detection, and evaluating some of the best state of the art methods: EAST [13], Pixellink [14], CRAFT [12], and PMTD [15]. The methods are evaluated on three challenging datasets, including the COCO-Text V2.0 [16], using a consistent methodology that contains several important challenges in the scene text detection.

2 Literature Review

In this section, a brief literature review on deep learning-based scene text detection techniques is presented. Table 1 offers a comparison among some of the recent state of the art text detection methods.

2.1 Regression-based Text Detection

Several methods [13, 28] adopted a general object detection regression-based (RB) framework, such as SSD [7] or Faster R-CNN [9], for text detection. They regard text regions as objects and predict candidate bounding boxes for text regions directly. For example, TextBoxes [28] modified the single-shot descriptor (SSD) [7] kernels by applying long default anchors and filters to handle the significant variation of aspect ratios of text instances to detect the various type of text shapes. Unlike TextBoxes, deep matching prior network (DMPNet) in [29] introduced quadrilateral sliding windows to handle detecting text under multiple orientations. There are many regression-based methods [22, 24] that have tried to solve the detection challenges of rotated and arbitrarily shaped text; for instance EAST [13] proposed a fast and accurate text detector, which makes dense predictions processed using locality-aware Non-Maximum Suppression (NMS) to detect multi-oriented text in an image without using manually designed anchors. Liao *et al.* [24] extended TextBoxes to TextBoxes++ by improving the network structure and the training process. Textboxes++ replaced the rectangle bounding boxes of text to quadrilateral to detect arbitrary-oriented text. RB methods usually have a simple post-processing framework to handle multi-oriented text. However, due to structural limitations in these methods it is not easy to represent accurate bounding boxes for arbitrary text shapes.

2.2 Segmentation-based Text Detection

Segmentation-based (SB) methods [14, 17, 23] classify text regions at the pixel level, making it possible to do word-level or character-level detection. They usually modify a segmentation framework like FCN [10] and Mask R-CNN [11]; for example, Zhang *et al.* [17] adopted FCN to predict the salient map of text regions, and TextSnake [23] adopts FCN as a base detector and extract text instances by detecting and assembling local components.

The preceding methods are trained to detect words in images, however it is challenging to use words as the basic unit for scene text detection as individual text characters may be represented in arbitrary shapes, therefore some recent text detection methods have trained deep learning models to detect text at the character-level [12, 18, 19]. For example, in [18] a saliency map of text regions, given by a dedicated segmentation network, uses character-level annotations to generate multi-oriented text bounding boxes. Later, Seglink [19] was trained to search for small text elements (segments) in the image and to link these segments to create word boxes using an additional post-processing step. Recently, CRAFT [12] used a weakly-supervised framework to detect individual characters in arbitrarily shaped text, which enables it to achieve the state of the art on benchmark datasets. Because text may appear in arbitrary shapes, recent methods usually adopted a segmentation framework as their backbone architecture, outperforming regression based methods in terms of multi-oriented text in several benchmark datasets. However, these types of methods require complex and time-consuming post-processing steps to produce the final detection result.

2.3 Hybrid Methods

Hybrid methods [15, 25, 26] use a combination of both segmentation and regression-based approaches for improving the perfor-

Table 1: Selected recent deep learning-based text detection methods.

Method	Model	Category			Architecture	Detection Target	Text-shape		Code
		RB	SB	Hy			MOT	CT	
Zhang *et al.* [17]	MOTD	–	✓	–	FCN	W	✓	–	–
Yao *et al.* [18]	STDH	–	✓	–	FCN	W	✓	–	✓
Shi *et al.* [19]	SegLink	✓	–	–	SSD	C,W	✓	–	–
He *et al.* [20]	SSTD	–	–	✓	SSD	W	✓	–	–
Hu *et al.* [21]	Wordsup	–	✓	–	FCN	C	✓	–	–
Zhou *et al.** [13]	EAST	✓	–	–	FCN	W,T	✓	–	✓
Ma *et al.* [22]	RRPN	✓	–	–	Faster R-CNN	W	✓	–	–
Long *et al.* [23]	TextSnake	–	✓	–	U-Net	W	✓	✓	✓
Liao *et al.* [24]	TextBoxes++	✓	–	–	SSD	W	✓	–	✓
Deng *et al.** [14]	Pixellink	✓	–	–	FCN	W	✓	–	✓
Liao *et al.* [25]	RRD	–	–	✓	SSD	W	✓	–	–
Lyu *et al.* [26]	MOSTD	–	–	✓	FCN	W	✓	–	–
Baek *et al.**[12]	CRAFT	–	✓	–	U-Net	C,W	✓	✓	✓
Liu *et al.** [15]	PMTD	–	–	✓	Mask-RCNN	W	✓	✓	✓
Liu *et al.* [27]	MB	–	✓	–	Mask-RCNN	W	✓	✓	✓

Note: *Methods have been considered for evaluation, W: Word, T: Text-line, C: Character, Hy:Hybrid Methods, MOT: Multi-Oriented Text, CT:Curved Text.

Table 2: Text detection datasets used for evaluation in this paper.

Dataset	Language	Year	# Images			# Text instance			Text Shape			Annotation level	
			Total	Train	Test	Total	Train	Test	H	CT	MO	Char	Word
ICDAR2013 [30]	EN	2013	462	229	233	1944	849	1095	✓	–	–	–	✓
ICDAR2015 [31]	EN	2015	1500	1000	500	17548	122318	5230	✓	✓	✓	–	✓
COCO-Text [32]	EN	2014	63686	43686	20000	145859	118309	27550	✓	✓	✓	–	✓
COCO-Text V2.0 [32]	EN	2014	63686	43686	20000	239506	–	–	✓	✓	✓	–	✓

Note: H: Horizontal, MO: Multi-Oriented, CT: Curved Text, EN: English.

mance of scene text detection, benefiting from the simple post-processing pipeline of regression-based methods, and the arbitrary-shape detection ability of segmentation based methods. For example, Lyu *et al.* [26] presented a method that can handle considerable variations in aspect ratio by grouping corner points to generate text boxes. Recently, based on the Mask R-CNN framework, Liu *et al.* [15] proposed the pyramid mask text detector (PMTD) for scene text detection, which achieved the state of the art performance in benchmark datasets [30, 31, 33]. PMTD assigns a soft pyramid label, i.e., a real value between 0 and 1, for each pixel in a text instance, and then reinterprets the obtained 2D soft mask into 3D space. However, hybrid methods have a complex framework and require more time for training compared to SB and RB approaches.

3 Experimental Results and Discussion

In this section, an experimental evaluation for a selected number of state-of-the-art methods [12–15] is conducted that covers the categories in Section 2. Table 2 shows the benchmark datasets that have been used for this evaluation. The datasets are described as follows:

ICDAR2013: This dataset [30] includes word-level annotations using rectangular boxes, which contains 229 and 233 images for training and testing, respectively. Most of the text instances of this dataset are horizontal and high-resolution.

ICDAR2015: This dataset [31] contains 1000 images for training and 500 images for testing. The annotations of this dataset are at the word-level represented using quadrilateral boxes. This dataset is more challenging in terms of orientation, illumination variation and complex background of text instances than ICDAR13 [30]. Most of the images in this dataset are from indoors environment.

COCO-Text: As shown in Table 2, when comparing this dataset [32] to the previous ones, this is the largest and the most challenging text detection dataset, which consists of 43,686 training images and 10,000 validation images [16, 18]. As in ICDAR13, the text images in this dataset are annotated in a word-level using rectangle bounding boxes. The text instances of this dataset are captured from different scenes: outdoor, sports field, grocery stores, etc. In this paper, we use the second version of this dataset, i.e., COCO-Text V2.0[1], for evaluation of detection approaches, as it contains 239,506 annotated text instances within 63,686 images.

Evaluation Metrics: For quantitative evaluation, we use the ICDAR15 IoU Metric [31], which is obtained for the jth ground-truth and ith detection bounding box as follow:

$$IOU = \frac{Area(G_j \cap D_i)}{Area(G_j \cup D_i)} \quad (1)$$

where a threshold of $IOU \geq 0.5$ is used for counting a correct detection for calculating the precision and recall. As in [12–15], we also use the H-mean (F-score) that is a function in the precision (P) and recall (R), and it is defined as follow:

$$\text{H-mean} = 2\frac{P \times R}{P + R} \quad (2)$$

For evaluation of scene text detection, recent deep learning-based methods, consisting of PMTD[2] [15] , CRAFT[3] [12], EAST[4] [13], and Pixellink[5] [14] have been selected. For an unbiased evaluation of existing approaches, we used the pre-trained models on ICDAR15 [31] directly from the authors' GitHub pages.

Generalization: One of the important characteristics of a scene text detector is to be generalizable, which shows how a trained model on one dataset is capable of detecting challenging text on other datasets. This evaluation strategy is an attempt to close the gap in evaluating text detection methods that are used to mainly trained and evaluated on a specific dataset. Therefore, to evaluate the generalization ability for the methods under consideration, we not only compare the detection performance of each model on ICDAR15 [31], which its training subset has been used for training the models but also on ICDAR13 [30] and COCO-Text v2.0 [16] test and validation subsets, respectively.

[1] https://bgshih.github.io/cocotext/#h2-explorer
[2] https://github.com/jjprincess/PMTD
[3] https://github.com/clovaai/CRAFT-pytorch
[4] https://github.com/ZJULearning/pixel_link
[5] https://github.com/argman/EAST

Challenges: One shortcoming of scene text detection datasets is that the challenges of each text instance in images are not labeled. To address this issue, we conduct an experiment to compare each method on some of these challenges.

Detection Precision: We would like to study how the detectors perform while increasing the IOU threshold for counting a true positive detection. Thus, we compute the H-mean at IOU thresholds between 0 and 1 to provide an evaluation of how a method is accurate under various constraints.

3.1 Quantitative Results

To evaluate the generalization ability of the methods, we compare the detection performance on ICDAR13 [30], ICDAR15 [31] and COCO-Text v2.0 [16] datasets. Table 3 illustrates the detection performance of the methods in [12–15]. Although the ICDAR13 is less challenging compared to ICDAR15, the detection performance of all methods decreased on ICDAR13 dataset that have not been used during training these methods. PMTD [15] had the minimum decline of about 0.6%, and Pixellink [14] that was the second-best methods in ICDAR15 had a maximum decrease of approximately 20%. However, all methods experienced a significant decrease in detection performance on COCO-Text dataset, which shows these models do not yet provide a generalization capability on a challenging dataset without being fine-tuned on part of the same dataset.

3.2 Qualitative Results

Figure 1 illustrates the detection performance for the studied methods on some challenging samples from ICDAR13, ICDAR15, and COCO-Text datasets. The detection results illustrate that the performance of all methods is far from perfect under challenging cases like difficult fonts, illumination variation, in-plane rotation, and low contrast text instances, especially when we have a combination of challenges affecting text instances. However, Hybrid regression and segmentation based methods, like PMTD [15], achieved the best H-mean values on all the three datasets, as they were able to handle better multi-oriented text, and methods that detect text at the character level, as in CRAFT [12], can perform better in detecting irregular shape text.

3.3 Discussion

To compare the precision of detection for each method, Figure 2 shows the H-mean computed at $0 \leq IOU \leq 1$. Overall, (1) decreasing the IOU threshold below 0.5 has a little effect on H-mean values of the detectors when evaluated on ICDAR13 and ICDAR15 datasets, but for COCO-Text dataset, due to the more difficult text instances in this dataset (Figure 2c) H-mean values are almost saturated for IOU < 0.4. (2) increasing the IOU-threshold over 0.6 results in rapidly reducing H-mean values offered by the detectors for all three datasets, which means the detected bounding box is not suitable, especially for text recognition task that usually requires an accurate localized text.

More specifically, for ICDAR13 (Figure 2a) the detection performance of all methods started from a relatively high H-mean similar to ICDAR15 (Figure 2b), because this dataset contains more horizontal and high-resolution text instances that are less challenging compared to that of ICDAR15 dataset. However, all methods experienced different H-mean curves in this dataset by increasing the IOU-threshold. For example, ICDAR13 dataset (Figure 2a) EAST [13] detector outperforms the PMTD [15] for IOU-thresholds > 0.8; because EAST detector uses a scale-invariant property that allows detecting more accurately text instances at different scales that are abundant in ICDAR13 dataset. Further, Pixellink [14] that ranked second on ICDAR15 has the worst detection performance on ICDAR13. This poor performance is also can be seen in challenging cases of the qualitative results in Figure 1.

The COCO-Text V2.0 [16] dataset is a good example of studying the generalization and precision performance of each method under adverse situations as it has more samples of text captured in the wild. Overall, as shown also in Table 3, all methods offer poor H-mean performance on this dataset (Figure 2c). For example, PMTD and CRAFT shown better performance than [13, 14] for IOU< 0.7. Since CRAFT is character-based methods, it performed better in localizing difficult-font words with individual characters. However, it is not robust to large-scale text due to the single-scale property of it. In addition, generally, the H-means of the detectors are declined to the half, from $\sim$ 60% to below of $\sim$ 30%, for IOU-threshold ≥ 0.7. We can also see this poor performance in the selected sample challenging cases in Figure 1, especially in the first row of the mentioned dataset.

4 Conclusion

Most recent deep learning-based methods have used multiple datasets and various evaluation metrics, which make the comparisons among the reported results difficult. In this paper, we experimented with comparing the performance of state-of-the-art scene text detection methods under adverse situations. By applying the pre-trained model provided by researchers, we showed that these methods do not offer the generalization capability on unseen datasets, and there are several challenges in the wild images, like in-plane-rotation, illumination reflection, partial occlusion, complex font styles, and perspective distortion, which most of the studied methods performed poorly. This study highlights also on the importance for having more descriptive annotations for text instances to allow future detectors to be trained and evaluated against more challenging conditions.

Acknowledgment

We would like to thank the Ontario Centres of Excellence (OCE), the Natural Sciences and Engineering Research Council of Canada (NSERC), and ATS Automation Tooling Systems Inc., Cambridge, ON, Canada for supporting this research work.

References

[1] H. Lin, P. Yang, and F. Zhang, "Review of scene text detection and recognition," *Arch. of Comput. Methods in Eng.*, pp. 1–22, 2019.

[2] S. B. Ahmed, M. I. Razzak, and R. Yusof, *Cursive Script Text Recognition in Natural Scene Images: Arabic Text Complexities*. Springer Nature, 2019.

[3] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," *TPAMI*, vol. 37, no. 7, pp. 1480–1500, 2015.

[4] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," *CoRR*, vol. abs/1811.04256, 2018.

[5] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Proc. Asian Conf. on Comp. Vision*. Springer, 2010, pp. 770–783.

[6] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *CVPR*, 2010, pp. 2963–2970.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *ECCV*, 2016, pp. 21–37.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016, pp. 779–788.

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

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

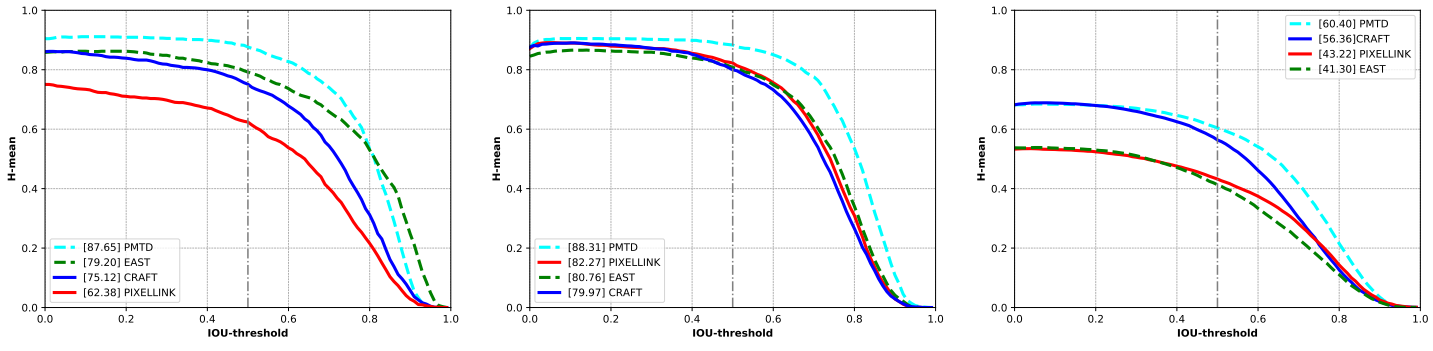
[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *CVPR*, 2017, pp. 2961–2969.

[12] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *CVPR*, 2019.

Table 3: Quantitative comparison among some of the recent text detection methods on ICDAR13, ICDAR15 and COCO-Text datasets, where ST: SyntheticText, IC15: ICDAR15, and IOU $\geq$ 0.5 is used, best and second best methods are highlighted by bold and underscore, respectively.

Method	Training Dataset(s)	ICDAR13			ICDAR15			COCO-Text			FPS
		P	R	H	P	R	H	P	R	H	
EAST [13]	ST+IC15	<u>84.86</u>	74.24	<u>79.20</u>	<u>84.64</u>	77.22	80.76	55.48	32.89	41.30	2.65
Pixellink [14]	IC15	62.21	62.55	62.38	82.89	<u>81.65</u>	<u>82.27</u>	<u>61.08</u>	33.45	43.22	4.88
CRAFT [12]	ST+IC15	72.77	<u>77.62</u>	75.12	82.20	77.85	79.97	56.73	<u>55.99</u>	<u>56.36</u>	<u>6.18</u>
PMTD [15]	IC17+IC15	**92.49**	**83.29**	**87.65**	**92.37**	**84.59**	**88.31**	**61.37**	**59.46**	**60.40**	**9.48**

Fig. 1: Sample qualitative comparison among CRAFT [12], PMTD [15], PixelLink [14], and EAST [13] on some Challenging examples from ICDAR13, ICDAR15 and COCO-text datasets. PO: partial occlusion, DF: difficult fonts, LC: Low Contrast, IV: illumination Variation, IB: Image blurriness, LR: Low Resolution, PD: perspective distortion, IPR: in-plane-rotation, and OT: Oriented Text, CT: Curved Text.



(a) ICDAR13 (b) ICDAR15 (c) COCO-Text

Fig. 2: Evaluation of the text detection performance for CRAFT [12], PMTD [15], PixelLink [14], and EAST [13] using H-mean versus IOU-threshold computed on (a) ICDAR13 [30], (b) ICDAR15 [31], and (c) COCO-Text [16] datasets.

[13] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "East: an efficient and accurate scene text detector," in *CVPR*, 2017, pp. 5551–5560.

[14] D. Deng, H. Liu, X. Li, and D. Cai, "Pixellink: Detecting scene text via instance segmentation," in *Proc. AAAI Conf. on Artif. Intell.*, 2018.

[15] J. Liu, X. Liu, J. Sheng, D. Liang, X. Li, and Q. Liu, "Pyramid mask text detector," *CoRR*, vol. abs/1903.11800, 2019.

[16] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," *arXiv preprint arXiv:1601.07140*, 2016.

[17] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," in *CVPR*, 2016, pp. 4159–4167.

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

[19] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *CVPR*, 2017, pp. 2550–2558.

[20] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in *CVPR*, 2017, pp. 3047–3055.

[21] H. Hu, C. Zhang, Y. Luo, Y. Wang, J. Han, and E. Ding, "Wordsup: Exploiting word annotations for character based text detection," in *CVPR*, 2017, pp. 4940–4949.

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

[23] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "Textsnake: A flexible representation for detecting text of arbitrary shapes," in *ECCV*, 2018, pp. 20–36.

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

[25] M. Liao, Z. Zhu, B. Shi, G.-s. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *CVPR*, 2018, pp. 5909–5918.

[26] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in *CVPR*, 2018, pp. 7553–7563.

[27] Y. Liu, S. Zhang, L. Jin, L. Xie, Y. Wu, and Z. Wang, "Omnidirectional scene text detection with sequential-free box discretization," *arXiv preprint arXiv:1906.02371*, 2019.

[28] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," in *Proc. AAAI Conf. on Artif. Intell.*, 2017.

[29] Y. Liu and L. Jin, "Deep matching prior network: Toward tighter multi-oriented text detection," in *CVPR*, 2017, pp. 1962–1969.

[30] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. De Las Heras, "Icdar 2013 robust reading competition," in *Proc. ICDAR*, 2013, pp. 1484–1493.

[31] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu *et al.*, "Icdar 2015 competition on robust reading," in *Proc. ICDAR*, 2015, pp. 1156–1160.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[33] M. Iwamura, N. Morimoto, K. Tainaka, D. Bazazian, L. Gomez, and D. Karatzas, "Icdar2017 robust reading challenge on omnidirectional video," in *Proc. ICDAR*, vol. 1, 2017, pp. 1448–1453.