Robust Scene Text Detection under Occlusion via Multi-Scale Adaptive Deep Network

My-Tham Dinh, Minh-Trieu Tran[0000−0002−5015−5604], Quang-Vinh Dang, and Guee-Sang Lee[0000−0002−8756−1382]

Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, South Korea
thamdinh.dmt@gmail.com, tmtvaa@gmail.com, quangvinh242003@yahoo.com, gslee@jnu.ac.kr

Abstract. Detecting text under occlusion in natural images is a challenge in scene text detection, which is severely sensitive and dramatically affects the performance of this field. Although some papers mention the solutions for the missing text problem, i.e., occlusion, they still fail on word regions with text bounding boxes splitting by the occlusion phenomena. In this paper, we first exploit the salient attention maps from Gradient Class Activation Maps Plus Plus (Grad-CAM++) on ImageNet to obtain knowledge of the important regions in the images. Moreover, to capture the diversity sizes of text instances and robustly enrich feature representations, we create a MulTi-scale adaptive Deep network (MTD). In addition to this task, from ICDAR 2015 benchmark, we build occluded text, namely Realistic Occluded Text Detection dataset (ROTD), and then combine a part of this new dataset with the ICDAR 2015 dataset for the training process to capture occluded text perception. Through these works, our model significantly improves the accuracy of text detection containing partially occluded text in natural scenes. Our proposed method achieves state-of-the-art results on partial occlusion text detection with $F1-score$ of 69.6% on ISTD-OC, 78.7% on our ROTD, and validates competitive performance $F1-score$ of 82.4% on ICDAR 2015 benchmark.

Keywords: Scene Text Detection · Multi-Scale Adapter Network · Grad-CAM++ · Occluded Text · Deep Learning.

1 Introduction

With the development of deep learning, scene text detection [1, 3], scene text segmentation [27, 29, 32], scene text recognition [31], and text spotting [30] have many achievements in scene text reading. As a key prior component of this field, text detection in natural scenes has played an essential role in computer vision, signal, and image processing. However, due to the variety of orientations, shapes, or sizes, it is still a challenging task, although many existing methods have achieved noticeable breakthroughs [1–3]. For example, DBNet++ [2] achieves consistently state-of-the-art accuracy and speed on five benchmarks of scene

Fig. 1. Several examples of failure text detection with text bounding boxes splitting by the occlusion phenomena of previous deep network architectures (a), and solved by our method (b) on ISTD-OC dataset.

text detection. Furthermore, TextPMs [3] obtains state-of-the-art performance in terms of detection accuracy both on polygonal and on quadrilateral datasets.

Unlike the previous prevalent problems in scene text detection, few researchers work on addressing partially occluded text problems [7, 8], which can significantly affect detection performance. For instance, [7] is detection and recognition task that can also achieve effective detection by restoring missing text. However, this method only assumes to detect text with few character-based distortion. Besides, in [8], the main task is to create ISTD-OC text occlusion dataset, involving different occlusion levels (from 0% to 100%), and evaluates the efficiency of state-of-the-art deep learning frameworks on ISTD-OC. Nevertheless, these frameworks are sensitive to occlusions and fail on text regions detection with text bounding boxes splitting by the occlusion phenomena, as in Figure 1.

In this paper, we design an approach to address this problem more efficiently. We apply a transfer learning Guided Grad-CAM++ Attention maps relying on Grad-CAM++ pre-trained on ImageNet [15] to obtain salient text regions. In addition, we give more robust feature representations with various sizes by exploring MTD. Our core contributions are as follows:

- We take advantage of the pre-trained from Grad-CAM++ from ImageNet [15] to gain the Guided Attention salient maps as one kind of specific information for training process, after that, we transfer the attention knowledge to our text detection under occluded text task.
- Our model MTD enhances both receptive fields by obtaining diverse scales and feature representation capability by learning multi-level information of features. Additionally, we adopt CBAM attention to improve the channel and spatial awareness abilities. Hence, our method is able to get richer feature representations.
- We also build our own occluded dataset ROTD based on ICDAR 2015 benchmark and combine only 10% as experimental results in Figure 6 with the

original one during training phase to learn more efficiently and accurately about occluded text awareness.

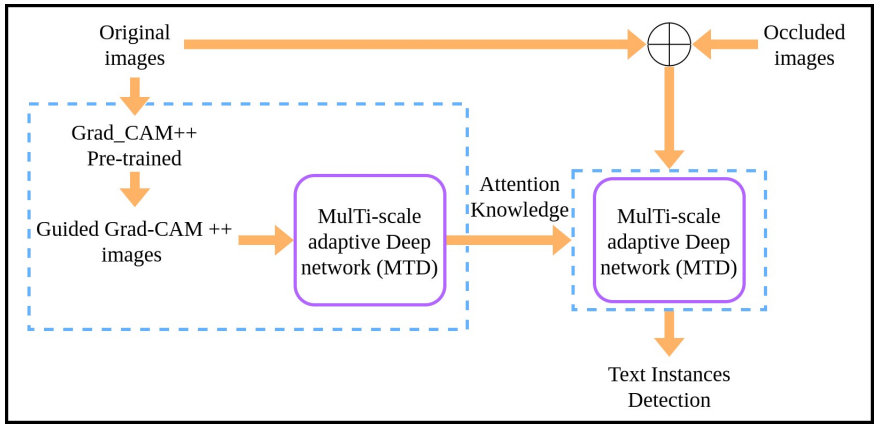

Fig. 2. This figure illustrates the overall architecture of the proposed method. Our approach includes three steps: Firstly, applying Guided Grad-CAM++ Attention maps for training process, next, transferring the learning knowledge to the main task, and finally, predicting text instances detection under occlusion.

2 Related Work

2.1 Scene Text Detection

Text detectors in natural images have achieved many remarkable results by many methods. Most of them are roughly divided into two phases, regression-based, and segmentation-based.

In the first category, several impressed research employed regression-based method [3, 17] that regresses directly bounding boxes of the text instances. EAST [13] could predict score maps from the fully convolutional network and multi-oriented text instances. Similarly, Deep-Reg [18] designed a per pixel-regression approach to detect multi-oriented tasks. However, it can be noted that these models are inadequate to cope with the occluded text challenge.

Additionally, another attractive method is segmentation-based [1, 11] that usually locates text regions following pixel-level prediction with post-processing algorithms. A progressive scale expansion algorithm is exploited in PSENet [11] to expand the detection areas with whole text instances. PAN [1] detected scene text instances and tackled overlap problem by clustering and aggregating text pixels by predicted similarity vectors.

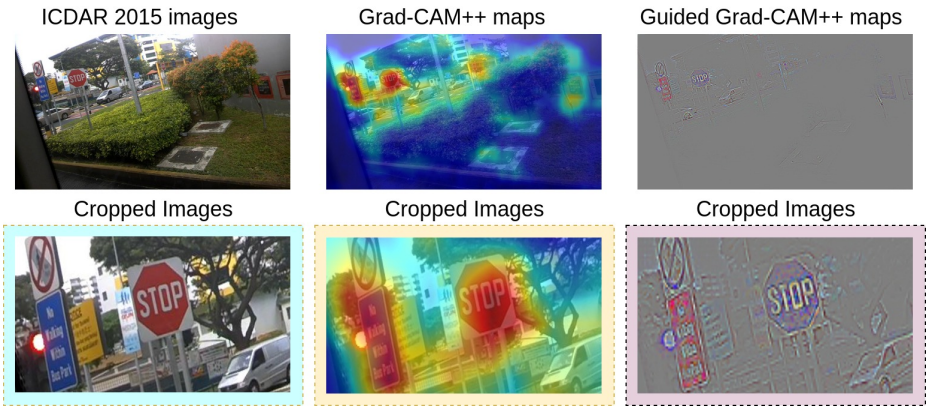

Fig. 3. Visualization of attention maps: Grad-CAM++ and Guided Grad-CAM++ from ICDAR 2015 images.

2.2 Occluded Text Detection

Partially occluded text in scene images, which may threaten prediction accuracy, is still a difficult challenge in scene text detection. [8] proved that several models from PAN [1], PSENet [11], CRAFT [12], and EAST [13] are still ineffective in detecting occluded text instances. In the same way, [7] handled the missing text issue by inheriting the strength of the characteristic Discrete Cosine Transform. Nevertheless, these methods have still failed significantly on text instances with text bounding boxes splitting by the occlusion phenomena. Therefore, this paper refines the capability of occlusion perception for scene text detection.

3 Methodology

Our overall architecture is illustrated in Figure 2. Firstly, we create Guided Grad-CAM++ maps from Grad-CAM++'s pre-trained on ImageNet [26] as in Figure 3, which focus on attention information and mitigate complex backgrounds. By learning those attention information in scene images, our approach can capture the spatial context of text instances. And then, passing them over MTD, including a ResNet18 [22] backbone, CBAM attention, a Multi-scale FEN (MFEN) [14], and features fusion. Due to the robustness of learning both extracted features with different scales and multi-level information features with less computation, our model enlarges the receptive fields and enhances the feature representation capabilities. After that, we follow as a PAN post-processing [1] of prediction network to detect bounding boxes of text instances as in Figure 4. Finally, transferring the learning attention knowledge of Guided Grad-CAM++ maps to the main task: Text detection containing occluded text. To help our model understand apparently occlusion knowledge from occluded images, we also employ a novel realistic occluded dataset (ROTD) by the OpenCV tool in Algorithm 1,

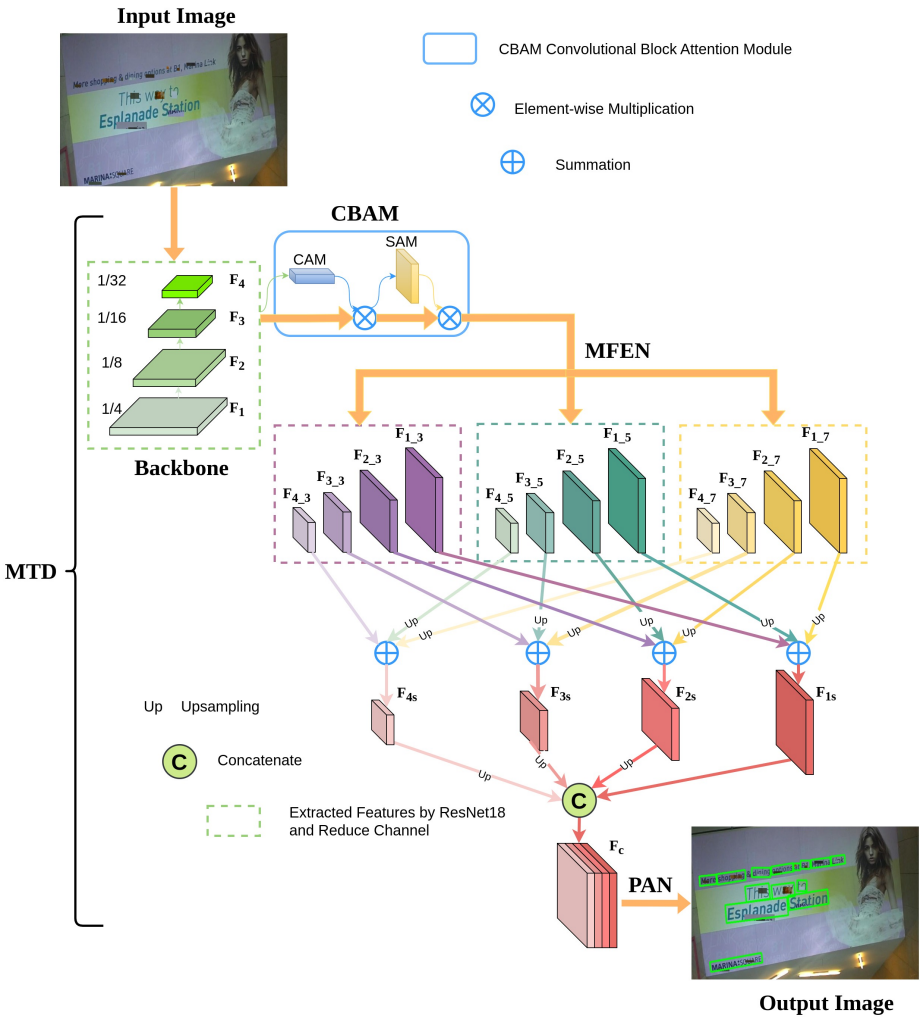

Fig. 4. Illustration of our proposed MTD network for occluded scene text detection.

and fuse a few number of ROTD images combined with the original ICDAR 2015 benchmark for training in the main phase.

3.1 Transfer Learning Knowledge from Guided Grad-CAM++ Attention

Gradient Class Activation Maps Plus Plus (Grad-CAM++) [15] improves the localization of both single, multiple instances, and produced perfect results for object classification and localization in the current state-of-the-art methods. Guided Grad-CAM++ Attention is carried out pointwise multiplication by

salient maps of Grad-CAM++ with pixel-space visualization by Guided Backpropagation. Therefore, to learn the important context in features, we apply pre-trained Grad-CAM++ on ImageNet [26] to obtain those salient maps containing the attention knowledge, then through them to MTD to get the valuable knowledge on ICDAR 2015. After that, transfer these weights to our main task: scene text detection under occlusion in scene images.

3.2 Multi-Scale Adaptive Deep Network

As illustrated in Figure 2, the input image (736x736) is fed into the feature extraction by ResNet18 [22] with pixel ratios 1/4, 1/8, 1/16, 1/32 corresponding F_1, F_2, F_3, and F_4. To reduce the time-consuming while keeping the general feature information, as PAN [1], we reduce the number of channels of each feature map to 128 by convolutional kernel 1x1. However, due to a lightweight backbone, features are often weak representation capabilities, so we apply CBAM attention [16], which perfectly illustrates the effectiveness of capturing spatial attention (SAM) along with channel attention (CAM). Thus, we can obtain richer feature representations. Then, MFEN is capable of the receptive field enhancement and different resolutions of text regions perception due to simultaneously progress with three different scales convolution kernels 3x3, 5x5, and 7x7. In the details, the structure of each MFEN is based on MobileNetv2 [23], which uses depthwise separable convolution (depthwise convolution 3x3 and pointwise convolution 1x1). Thus, the spatial information on features F_{1_n}, F_{2_n}, F_{3_n}, F_{4_n} (n is kernel size) are captured more adequately. Afterward, to prepare for predicting task, features of different depths are integrated into an enriched feature F_c by upsampling and concatenating extracted features. Finally, we inherited the post-processing of PAN [1] that detects text instances followed by pixel aggregation algorithms. This method clusters the neighbor pixels and merges them in the iterating process; consequently, text kernel is gradually expanded to text region.

$$F_{2\_n} = Upsample(F_{2\_n}|F_{1\_n}) \tag{1}$$

$$F_{3\_n} = Upsample(F_{3\_n}|F_{1\_n}) \tag{2}$$

$$F_{4\_n} = Upsample(F_{4\_n}|F_{1\_n}) \tag{3}$$

$$F_c = Concatenate((F_{1\_n}, F_{2\_n}, F_{3\_n}, F_{4\_n})|1) \tag{4}$$

3.3 Loss Function

Our loss function L can be formulated as a weighted sum of the loss for text region, text kernel and sum of loss for similarity vector by segmentation network:

$$L = L_{reg} + \alpha L_{ker} + \beta(L_{agg} + L_{disc}) \tag{5}$$

where L_{reg}, and L_{ker} define loss of text regions and text kernels as Eq. 6, Eq. 7, respectively. L_{agg}, and L_{disc} are aggregation loss and discrimination loss of

post-processing stage as in PAN in Eq. 8, Eq. 9. According to the numeric values of the losses, $\alpha = 0.5$, $\beta = 0.25$ are two constants selecting to keep the balance among these losses.

In more details, prediction of text regions and text kernels are basically a pixel-wise classification text or non-text problem, so we apply dice loss [24] to handle these works.

$$L_{reg} = \sum_{i} Dice(P_{reg}, G_{reg}) \tag{6}$$

$$L_{ker} = \sum_{i} Dice(P_{ker}, G_{ker}) \tag{7}$$

where P_{reg}, G_{reg} are the prediction and ground truth of text region, respectively. P_{ker}, G_{ker} are the prediction and ground truth of text kernel.

Besides, in post-processing, we adopt loss function from PAN by using aggregation loss L_{agg} and L_{dis} as shown below:

$$L_{agg} = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{|T_j|} \sum_{pix \in T_j} ln(D(pix, T_{ker_j}) + 1) \tag{8}$$

where N, T_j define the number and jth of text instances. The distance between text pixel pix and kernel jth T_{ker_j} of the same instance should be small, which is denoted by $D(pix, T_{ker_j})$. This function is calculated by maximum of similarity vector between pix and T_{ker}. This function is set with a constant 0.5 as PAN experimentally.

In addition to this, to reduce overlap among text regions, the text instance vectors should keep apparently discrimination. In training phase, PAN implemented this discrimination loss L_{disc} as below:

$$L_{disc} = \frac{1}{N_j N_k} \sum_{j=1}^{N} \sum_{k=1}^{N} \ln(D(T_{ker_j}, T_{ker_k}) + 1) \tag{9}$$

Similar to aggregation loss, where N, $D(T_{ker_j}, T_{ker_k})$ define the number of text instances, the distance between the text kernel T_{ker_j} and the text kernel T_{ker_k}, respectively, corresponding jth, kth.

4 Experimental Results

ICDAR 2015 [25] is the incidental scene text of challenge four on the website https://rrc.cvc.uab.es/?ch=4. It consists of 1000 incidental natural images for training process and 500 images for testing set. ICDAR 2015 dataset is one of the popular datasets for scene text detection, including word-level text instances with multi-oriented texts.

ISTD-OC [8], named Incidental Scene Text Dataset - Occlusion, was conducted for occluded text detection and recognition task in workshop CBDAR

Algorithm 1 Realistic Occluded Text Detection (ROTD) dataset

Input: ICDAR 2015 images
Method: OpenCV
Output: Occluded text images

while Bounding box (bbox) has described-text: **do**
  Find location of bbox with described-text (x_a, y_a).
  Get color value $(0 - 255)$ of a random pixel inside the bounding box above.
  if pixel=255 **then**
    pixel=0
  else
    pixel=pixel/255
  end if
  Initial: Set n, r, N
  Draw shape with initial parameters and the color value got from Step 1.
  Save occluded images.
end while

2021. ISTD-OC contains the rectangle shape for ICDAR 2015 benchmark with different levels of occlusions from zero to a hundred percent. The number of images is 1500 occluded images for detection, but only 500 evaluation images are published.

We create a novel Realistic Occluded Text Detection (ROTD) dataset in Algorithm 1 with two steps: The first, find the color of a random pixel inside the box, providing localization in Ground Truth of ICDAR 2015. In the second step, draw the arbitrary shape inside the bounding box with the color value obtained above. The arbitrary shape is initialized with the number of possibly sharp edges $n = 7$, the magnitude of the perturbation from unit circle $r = 0.7$, and the number of points in the path $N = n*3+1$, experimentally. The difference between our proposed dataset and ISTD-OC is incidental shape and color, making the part of missing texts look real as in Figure 5. As the number of ICDAR 2015 images, our proposed dataset includes 1000 training and 500 testing images.

In this work, to help model understand deeply context of texts and even occlusion texts, we choose random only 10% training images following our experimental results (the highest F1-score performance 67.21%) in Figure 6 and associate with original ICDAR 2015 training set during training stage, totally 1100 images (1000 original ICDAR 2015 images and 100 occluded text ROTD images). Additionally, to compare fairly with ISTD-OC, as proved in paper [8], we selected 70% occlusion on ISTD-OC as a standard testing set for our evaluation.

The comparison results with previous methods on occluded text ISTD-OC, ICDAR 2015 benchmark, and our own ROTD dataset are demonstrated in Table 1, Table 2, and Table 3, respectively. As shown, our method is superior for 70% occluded text detection on ISTD-OC by 69.6%, 78.7% on ROTD dataset, and

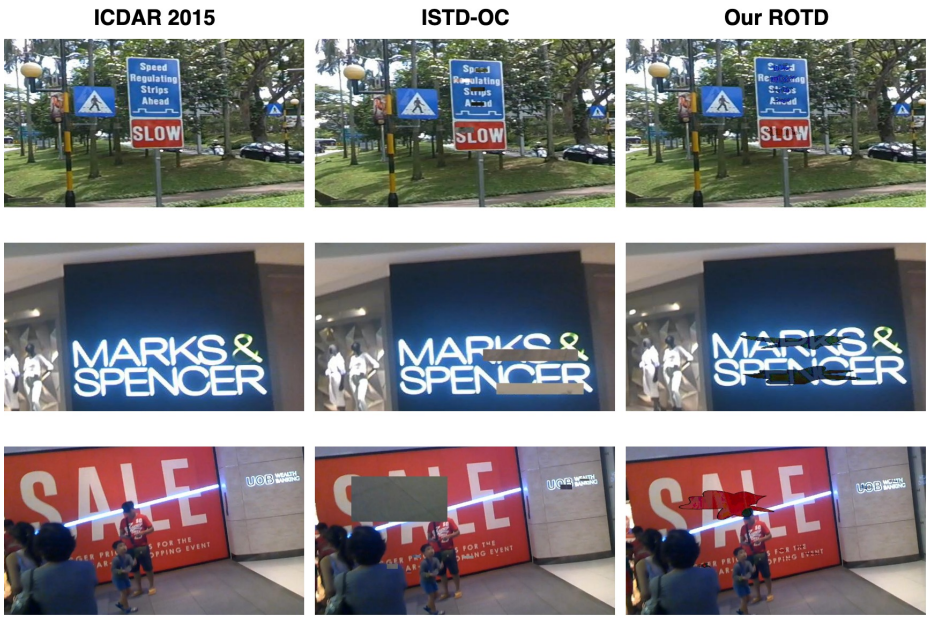

Fig. 5. ICDAR 2015 is represented for normal scene texts, ISTD-OC is occluded text images with rectangle shape, our ROTD is occluded text images with arbitrary shape.

performs better in detecting the text instances on ICDAR 2015 by 82.4%. Several visualizations are shown in Figure 7.

Table 1. Comparison of Occluded Text Detection on ISTD-OC. *≈ is represented the results from the mentioned graph performances in [8].

Method	Precision	Recall	F-score
PAN [1]	≈ 64	≈ 44	≈ 61
PSE-Net [11]	≈ 58	≈ 52	≈ 62
EAST [13]	≈ 43	≈ 51	≈ 60
PAN [1] (1100 images)	**78.5**	57.7	66.5
Ours (1100 images)	77.7	**63.0**	**69.6**

5 Conclusion

In this paper, we have presented a method for addressing text bounding boxes splitting by the occlusion phenomena problem with three stages: We first focus on perceiving attention information from Guided Grad-CAM++ maps to prepare knowledge for the main task. Next, we employ a novel MTD network, which

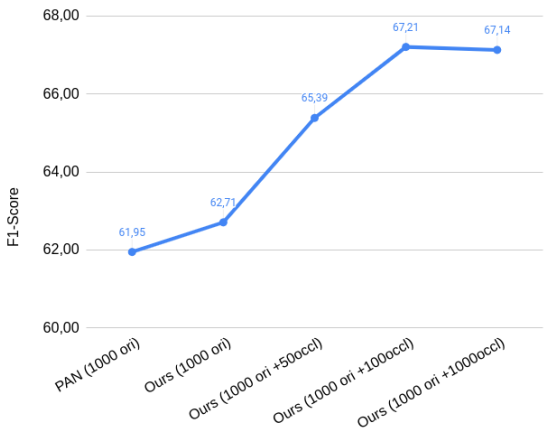

Fig. 6. This graph shows the comparison of F1-score performances of proportional scene text occlusion images.

Table 2. Comparison of state-of-the-art scene text detection on ICDAR 2015 without external data.

Method	Precision	Recall	F-score
PAN [1]	82.9	77.8	80.3
PSE-Net [11]	81.5	**79.7**	80.6
EAST [13]	83.6	73.5	78.2
MFEN [14]	84.5	**79.7**	82.0
Ours (1100 images)	**85.8**	79.3	**82.4**

Table 3. Comparison of our model with validation on ROTD dataset.

Method	Precision	Recall	F-score
PAN [1]	70.1	50.2	58.5
Ours (1000 images)	72.2	55.0	62.4
Ours (1100 images)	**82.1**	**75.6**	**78.7**

is capable of enlarging the receptive fields and enriching feature representations while bringing minor extra computation. Finally, to aware text occlusion knowledge, we conduct a new dataset ROTD, and combine a part of them with original images (ICDAR 2015) for training process. Therefore, the proposed method outperforms the state-of-the-art on ISTD-OC dataset. Extensive experiment on ICDAR 2015 shows the competitive result compare to other recent methods.

6 Acknowledgements

In the future, This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) & funded by the Korean government (MSIT) (NRF-2019M3E5D1A02067961) and by Basic

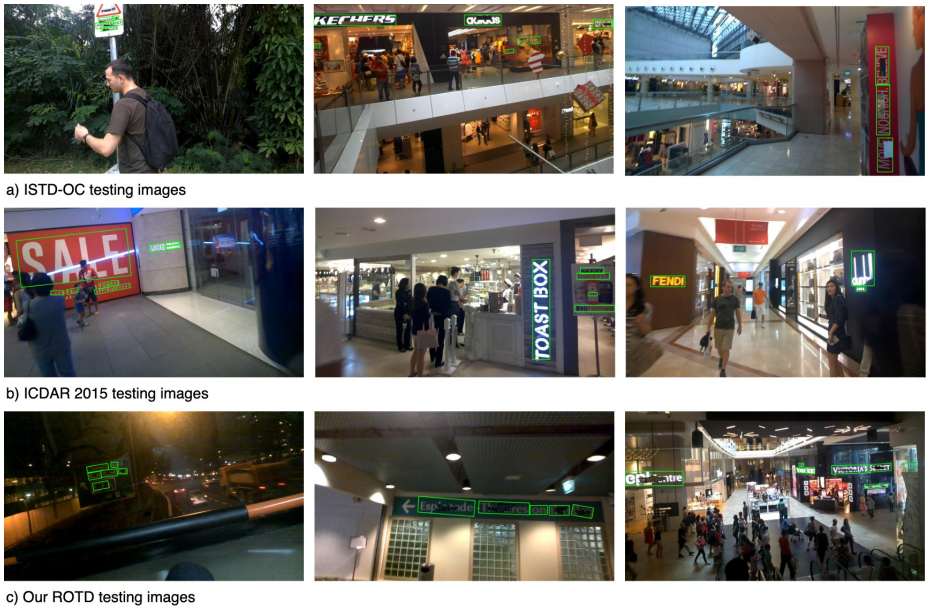

Fig. 7. Visual comparisons of bounding box representations for text detection on three sets: ISTD-OC, ICDAR 2015 and our ROTD.

Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2018R1D1A3B05049058 & NRF-2020R1A4A1019191).

References

1. Wang Wenhai, Xie Enze, Song Xiaoge, Zang Yuhang, Wang Wenjia, Lu Tong, Yu Gang, Shen Chunhua, "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8440–8449, 2019.
2. Liao Minghui, Zou Zhisheng, Wan Zhaoyi, Yao Cong, Bai Xiang, "Real-time scene text detection with differentiable binarization and adaptive scale fusion," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
3. Zhang Shi-Xue, Zhu Xiaobin, Chen Lei, Hou Jie-Bo, Yin Xu-Cheng, "Arbitrary Shape Text Detection via Segmentation with Probability Map," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
4. Tang Jingqun, Zhang Wenqing, Liu Hongye, Yang MingKun, Jiang Bo, Hu Guanglong, Bai Xiang, "Few Could Be Better Than All: Feature Sampling and Grouping for Scene Text Detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4563–4572, 2022.
5. Yin Xu-Cheng, Yin Xuwang, Huang Kaizhu, Hao Hong-Wei, "Robust text detection in natural scene images," in IEEE transactions on pattern analysis and machine intelligence, pp. 970–983, 2013.

6. Chen Zhe, Wang Wenhai, Xie Enze, Yang ZhiBo, Lu Tong, Luo Pin, "FAST: Searching for a Faster Arbitrarily-Shaped Text Detector with Minimalist Kernel Representation," in arXiv preprint arXiv:2111.02394, 2021.
7. Mittal Ayush, Shivakumara Palaiahnakote, Pal Umapada, Lu Tong, Blumenstein Michael, "A new method for detection and prediction of occluded text in natural scene images," in Signal Processing: Image Communication, pp. 116512, 2022.
8. Geovanna Soares Aline, Leite Dantas Bezerra Byron, Baptista Lima Estanislau, "How Far Deep Learning Systems for Text Detection and Recognition in Natural Scenes are Affected by Occlusion?," in International Conference on Document Analysis and Recognition, pp. 198–212, 2021.
9. Zhou Bolei, Khosla Aditya, Lapedriza Agata, Oliva Aude, Torralba Antonio, "Learning deep features for discriminative localization," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921–2929, 2016.
10. Mittal Ayush, Shivakumara Palaiahnakote, Pal Umapada, Lu Tong, Blumenstein Michael, Lopresti Daniel, "A new context-based method for restoring occluded text in natural scene images," in International Workshop on Document Analysis Systems, pp. 466–480, 2020.
11. Wang Wenhai, Xie Enze, Li Xiang, Hou Wenbo, Lu Tong, Yu Gang, Shao Shuai, "Shape robust text detection with progressive scale expansion network," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9336–9345, 2019.
12. Baek Youngmin, Lee Bado, Han Dongyoon, Yun Sangdoo, Lee Hwalsuk, "Character region awareness for text detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9365–9374, 2019.
13. Zhou Xinyu, Yao Cong, Wen He, Wang Yuzhi, Zhou Shuchang, He Weiran, Liang Jiajun, "East: an efficient and accurate scene text detector," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 5551–5560, 2017.
14. Dinh My-Tham and Lee Guee-Sang, "Arbitrary-shaped Scene Text Detection based on Multi-scale Feature Enhancement Network," in Korea Computer Congress, pp. 669–671, 2022.
15. Chattopadhay Aditya, Sarkar Anirban, Howlader Prantik, Balasubramanian Vineeth N , "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in 2018 IEEE winter conference on applications of computer vision (WACV), pp. 839–847.
16. Woo Sanghyun, Park Jongchan, Lee Joon-Young, Kweon In So, "Cbam: Convolutional block attention module," in Proceedings of the European conference on computer vision (ECCV), pp. 3–19, 2018.
17. Dai Pengwen, Zhang Sanyi, Zhang Hua, Cao Xiaochun, "Progressive contour regression for arbitrary-shape scene text detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7393–7402, 2021.
18. He Wenhao, Zhang Xu-Yao, Yin Fei, Liu Cheng-Lin, "Deep direct regression for multi-oriented scene text detection," in Proceedings of the IEEE international conference on computer vision, pp. 745–753, 2017.
19. Sheng Tao, Chen Jie, Lian Zhouhui, "Centripetaltext: An efficient text instance representation for scene text detection," in Advances in Neural Information Processing Systems, pp. 335–346, 2021.
20. Tian Zhuotao, Shu Michelle, Lyu Pengyuan, Li Ruiyu, Zhou Chao, Shen Xiaoyong, Jia Jiaya, "Learning shape-aware embedding for scene text detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4234–4243, 2019.

21. Selvaraju Ramprasaath R, Cogswell Michael, Das Abhishek, Vedantam Ramakrishna, Parikh Devi, Batra Dhruv, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE international conference on computer vision, pp. 618–626, 2017.
22. He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.
23. Howard Andrew G, Zhu Menglong, Chen Bo, Kalenichenko Dmitry, Wang Weijun, Weyand Tobias, Andreetto Marco, Adam Hartwig, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," in arXiv preprint arXiv:1704.04861, pp. 770–778, 2017.
24. Sudre Carole H, Li Wenqi, Vercauteren Tom, Ourselin Sebastien, Jorge Cardoso M, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 240–248, 2017.
25. Karatzas Dimosthenis, Gomez-Bigorda Lluis, Nicolaou Anguelos, Ghosh Suman, Bagdanov Andrew, Iwamura Masakazu, Matas Jiri, Neumann Lukas, Chandrasekhar Vijay Ramaseshan, Lu Shijian and others, "ICDAR 2015 competition on robust reading," in 2015 13th international conference on document analysis and recognition (ICDAR), pp. 1156–1160, 2015.
26. Deng Jia, Dong Wei, Socher Richard, Li Li-Jia, Li Kai, Fei-Fei Li, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255, 2009.
27. Dang Quang-Vinh, Lee Guee-Sang, "Document image binarization with stroke boundary feature guided network," in IEEE Access, pp. 36924–36936, 2021.
28. Wu Yonghui, Schuster Mike, Chen Zhifeng, Le Quoc V, Norouzi Mohammad, Macherey Wolfgang, Krikun Maxim, Cao Yuan, Gao Qin, Macherey Klaus and others, "Google's neural machine translation system: Bridging the gap between human and machine translation," in arXiv preprint arXiv:1609.08144, 2016.
29. Dang Quang-Vinh, Lee Guee-Sang, "Document Image Binarization by GAN with Unpaired Data Training," in International Journal of Contents, pp. 8–18, 2020.
30. Wang Wenhai, Xie Enze, Li Xiang, Liu Xuebo, Liang Ding, Yang Zhibo, Lu Tong, Shen Chunhuag, "Pan++: Towards efficient and accurate end-to-end spotting of arbitrarily-shaped text," in IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 5349–5367, 2021.
31. Aberdam Aviad, Litman Ron, Tsiper Shahar, Anschel Oron, Slossberg Ron, Mazor Shai, Manmatha R, Perona Pietrog, "Sequence-to-sequence contrastive learning for text recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15302–15312, 2021.
32. Xu Xingqian, Zhang Zhifei, Wang Zhaowen, Price Brian, Wang Zhonghao, Shi Humphrey, "Rethinking text segmentation: A novel dataset and a text-specific refinement approach," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12045–12055, 2021.
33. Deng Dan, Liu Haifeng, Li Xuelong, Cai Deng, "Pixellink: Detecting scene text via instance segmentation," in Proceedings of the AAAI Conference on Artificial Intelligence, 2018.