A Text-context-aware CNN Network for Multi-oriented and Multi-language Scene Text Detection

Yao Xiao[‡] Minglong Xue[‡] Tong Lu[‡⋆] Yirui Wu[†‡] Shivakumara Palaiahnakote[§]
[‡]National Key Lab for Novel Software Technology, Nanjing University
[†]College of Computer and Information, Hohai University
[§]Department of Computer System and Information Technology, University of Malaya
{iam_xiaoyao@126.com;xueml@smail.nju.edu.cn;lutong@nju.edu.cn;wuyirui@hhu.edu;shiva@um.edu.my}

***Abstract*—The existing deep learning based state-of-the-art scene text detection methods treat scene texts a type of general objects, or segment text regions directly. The latter category achieves remarkable detection results on arbitrary-orientation and large aspect ratios of scene texts based on instance segmentation algorithms. However, due to the lack of context information with consideration of scene text unique characteristics, directly applying instance segmentation to text detection task is prone to result in low accuracy, especially producing false positive detection results. To ease this problem, we propose a novel text-context-aware scene text detection CNN structure, which appropriately encodes channel and spatial attention information to construct context-aware and discriminative feature map for multi-oriented and multi-language text detection tasks. With high representation ability of text-context-aware feature map, the proposed instance segmentation based method can not only robustly detect multi-oriented and multi-language text from natural scene images, but also produce better text detection results by greatly reducing false positives. Experiments on ICDAR2015 and ICDAR2017-MLT datasets show that the proposed method has achieved superior performances in precision, recall and F-measure than most of the existing studies.**

***Keywords*-Text Context Aware Information; Attention Module; Scene Text Detection; Mask R-CNN;**

I. Introduction

Scene text detection is challenging due to the variations of texts in aspect ratios, scales, orientations, languages, extreme illumination, occlusion and complex backgrounds. Inspired by the thoughts of utilizing deep architectures for general object detection and semantic segmentation, researchers have proposed variety of models by regarding text as an instance of objects or segments, which greatly improves accuracy and robustness of text detection tasks especially facing variations of scene texts. Following such a trend, we category the recent deep learning based text detection methods as regression and segmentation based methods.

Regression based methods predict text bounding boxes by adopting object detection frameworks. For example, TextBoxes++ [1] extends the thought of a famous object detection model, i.e., SSD [2], to use relatively ”long” default boxes and ”long” convolutional filters, which successfully cope with the extreme aspect ratios of text instances. Inspired by the recent progress of fine-level image segmentation, segmentation based methods cast text detection as a semantic segmentation problem, thus functioning particularly well on rotated text components. For example, Zhang et al. [3] extract text blocks from a segmentation map computed by a Fully Convolutional Network (FCN), and then perform post-processing steps to compute text lines based on several priori rules.

To deal with texts of any orientation, shape and language, we follow the idea of a recent state-of-the-art instance segmentation framework, i.e., Mask R-CNN [4], to perform text detection with two parallel workflows which are horizontal text bounding boxes detection and text instance segmentation. We further introduce a novel Text-Context-Aware Attention Module (TCAM) to involve text context information for less false positive errors and thus better detection results. Specifically, the proposed TCAM extracts informative information from multi-scale receptive fields and enhances feature responses of saliency regions known as channel and spatial attentions, respectively.

The main contribution of this paper is to propose a text-context-aware network for accurate and robust scene text detection. This mehtod can detect both multi-oriented and multi-language texts in a unified manner. As far as we know, the proposed TCAM firstly involves both channel and spatial attention information to construct a more task-specified feature map for scene text detection. Besides, its simple structure and implementation help it easily transform to another text related tasks. Based on the feature map produced by TCAM, we specially perform accurate text instance segmentation to focus on dealing with texts of any shape and orientation in a natural and effective way. The proposed method achieves superior performances on two public datasets, which proves the effectiveness of TCAM and the following instance-level text mask segmentation.

II. Related Work

A. Scene Text Detection

There are two categories of deep learning based text detection approaches, i.e., regression based and segmentation based methods. Regression based methods view text

2379-2140/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDAR.2019.00116

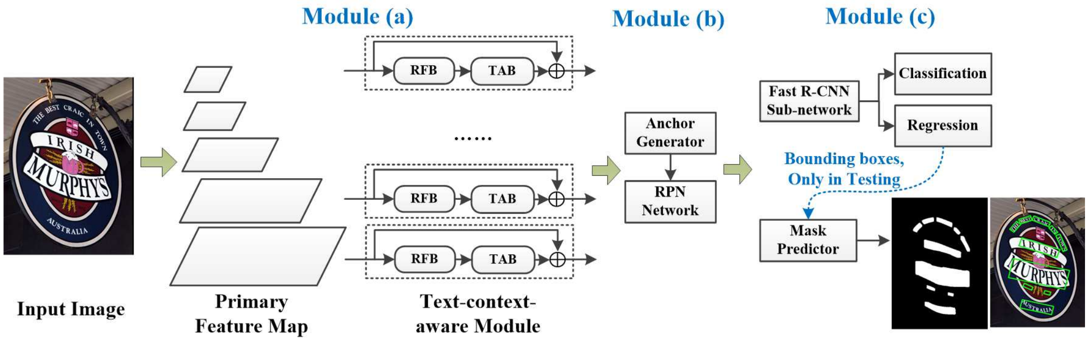

Figure 1. Network design of the proposed method, where Module (a) enhances multi-scale feature map by TCAM, Module (b) generates text region proposals for next step, and Module (c) performs text bounding boxes regression and mask prediction in parallel ways. It is noted that structure of the proposed TCAM is represented with a dotted rectangle.

detection as a special case of general object detection, and detect word or text line level bounding boxes directly. EAST [5] proposes a simple yet powerful pipeline that yields fast and accurate text detection for words or text lines of arbitrary orientations and quadrilateral shapes in full images with a single neural network. Busta *et al.* [6] build their end-to-end framework on the YOLO object detector [7], which simultaneously detects and recognizes texts in scene images. These methods are generally fast in speed due to less burden of complicated stages in training and testing. However, these methods are struggled to deal with curved or arbitrary texts.

Segmentation based methods regard text detection as a semantic segmentation problem or instance segmentation problem. To better separate adjacent text instances, Wu and Natarajan [8] distinguish each pixel by deep neural networks into three categories, that is, non-text, text border and text. Inspired by an instance segmentation framework FCIS [9], Dai et al. propose FTSN [10] to address scene text detection, which has the ability to detect both straight and curved texts. Recently, Xie et al. and Lyu et al. have achieved remarkable performance in curved text detection [11] and end-to-end text spotting [12] respectively by adopting Mask R-CNN [4]. In our paper, we also adopt the latest state-of-the-art instance segmentation framework, i.e., Mask R-CNN to pursue text detection with high accuracy and low computation cost.

B. Attention Model

Attention model, the selectively focusing mechanism, has demonstrated to be very effective in various applications. Liu et al. [15] propose a global context-aware attention LSTM for RGB-D action recognition, which recurrently optimize the global contextual information and further utilizes it as an informative function to assist accurate action recognition. However, former visual attention models are generally spatial, i.e., the attention is modeled as spatial probabilities that re-weight the last conv-layer feature map of a CNN encoding an input image. Chen et al. [13] thus introduce a novel convolutional neural network dubbed SCA-CNN that incorporates Spatial and Channel Attentions in a CNN for the task of image caption. However, their proposed CNN is specially and carefully designed, so that it's too complicated to transform into other tasks. Recently, Woo et al. proposed Convolutional Block Attention Module(CBAM) [14] to achieve excellent performance in several tasks by combining channel attention and spatial attention. Liu et al. proposed Receptive Field Block(RFB) [16] to enhance the feature representation of one-stage object detector SSD [2]. Our TCAM is inspired by CBAM [14] and RFB [16] to bring their advantages together as a single framework: TCAM.

III. The Proposed Method

We involve TCAM to enhance feature representation with text-context-aware characteristics. Based on text-context-aware feature map, we perform text box regression and instance-level text mask segmentation in a parallel way to improve accuracy. We organize this section by first illustrating the total network architecture, and then describing structure of TCAM in details.

A. Network Design

The network design of the proposed method is presented in Fig. 1, which consists of three modules, namely, (a) multi-scale feature map enhanced by TCAM, (b) text proposal generation, (c) a Fast R-CNN based sub-network to detect horizontal bounding boxes containing texts, and a Mask Predictor sub-network to predict text instance masks. The proposed method firstly constructs a pyramid structure with different size ratios to deal with the multi-scale difficulty of scene texts, where size ratio is defined as $\frac{1}{2^{l+1}}, l \in \{1, ..., 5\}$ and l refers to the level index of the pyramid. After constructing the pyramid, we utilize pre-trained ResNet-50 on the ImageNet dataset to extract primary feature map for different levels F_l, which are represented as $C \times H \times W$ in size. Afterwards, F_l is separatively enhanced by the proposed TCAM at different levels for more informative representation. Following the idea and procedures of Mask

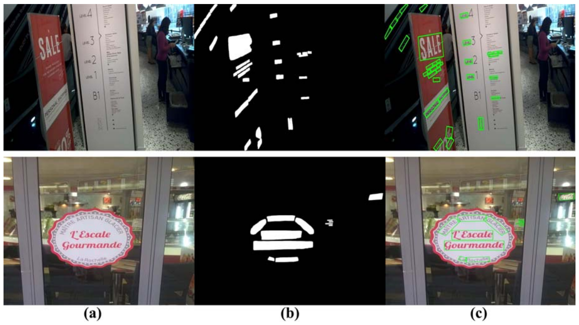

Figure 2. Workflow of the proposed method: (a) input images; (b) text instance segmentation results, and (c) the final processed quadrilateral results. The workflow shows our method can handle both multi-oriented and multi-language texts in a unified manner.

R-CNN [4], we design Module (b) and (c) to utilize its state-of-the-art performance on instance segmentation, which helps achieve desirable text detection results on multi-oriented and multi-language texts. During Module (b), a manually settled anchor generator is proposed to produce quantity of anchors, which are further refined by the RPN network as the input for Fast R-CNN sub-network and Mask Predictor, respectively. Finally, Module (c) classifies input refined anchors as text or non-text and regresses locations of bounding boxes. Meanwhile, another thread in Module (c) outputs extracted text masks to deal with the difficulties of multi-oriented and multi-language texts.

During training, there are several parts required to be optimized, such as RPN network, Fast R-CNN sub-network and Mask Predictor. To jointly training these parts, we involve their estimation as a multi-task loss, which is represented as

$$Loss = L_{rpn} + L_{cls} + \lambda L_{box} + L_{mask} \tag{1}$$

where L_{rpn}, L_{mask}, L_{cls} and L_{box} denote the loss of RPN network, Mask Predictor, classification and regression loss of Fast R-CNN, respectively, λ is a weight parameter and we define $\lambda = 3$ by experiments. L_{rpn}, L_{mask}, L_{cls} and L_{box} are identical as those defined in Mask R-CNN [4]. We optimize Equ. 1 in an end-to-end manner instead of alternative training for these three parts.

During test, resulting bounding boxes are firstly extracted from the masks predicted by Mask Predictor with an area filter. Afterwards, we perform Non-Maximum Suppression (NMS) with an IoU threshold 0.25 on extracted bounding boxes to generate the final results. It is noted that NMS is performed based on oriented rectangles or quadrangles, which differs from the usual design in object detection, where NMS is performed based on horizontal rectangles.

B. Text-Context-Aware Attention Module

Directly applying instance segmentation algorithms like Mask R-CNN on scene text detection may lead to low accuracy with quantity of false positive results due to the lack of involving text-aware context information to help distinguish between text regions and text-like backgrounds. We thus design TCAM, which performs channel and spatial attention operations to construct context-aware and discriminative feature map for text detection. We show the structure of TCAM in Fig. 1, where Receptive Field Block (RFB) is used to enhance feature representation for higher discrimination ability, while Text-Context-Aware Attention Block (TAB) describes how channel and spatial attentions are involved into feature map construction. However, such multi-stage process of TCAM and the following modules make it hard to train for convergency. To ease this difficulty, we propose a residual shortcut on TCAM, which makes gradient descent propagate in a much easier way. Therefore, the computing formula of TCAM can be represented as

$$T(F_l) = F_l \oplus A(R(F_l)) \tag{2}$$

where $\oplus$ implies element-wise addition, function $R()$ and $A()$ represent process of RFB and TAB, respectively.

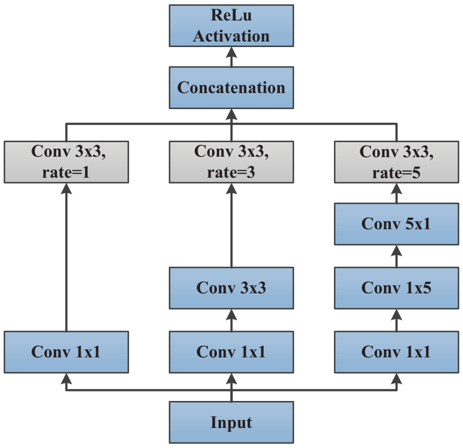

Figure 3. Architecture of RFB, where gray rectangles represent astrous convolution layers and ”rate” in these rectangle refers to the parameter for astrous rate we used in RFB.

Receptive Field Block Inspired by [16], we design a text-specified RFB to simulate the relationship between size and eccentricity of receptive fields in human visual systems. In other words, we aim to ensure that positions near the center have larger weights than faraway ones with variety of kernels. Essentially, the proposed RFB makes use of multi-branch convolutional layers with varying and text-specified kernel sizes, which is benefit to enhance the feature discrimination ability by introducing information of different receptive fields for unity representation.

The structure of RFB is shown in Fig. 3. It is noted that we use the combination of two kinds of convolutional layers, i.e., normal and atrous convolutional layers. Astrous convolution is designed to capture information of a larger area while keeping the number of parameters unchanged. More-

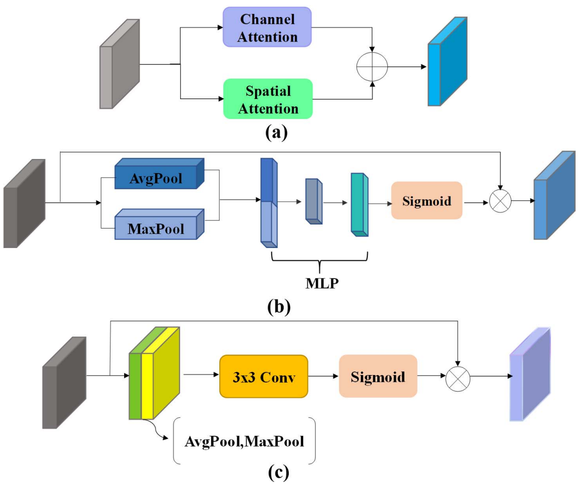

Figure 4. Architecture of TAB, where (a) implies that TAB consists of a channel and spatial attention scheme, (b) and (c) represent architectures of channel and spatial attention scheme, respectively.

over, we utilize convolutional layers with special chosen kernel sizes, i.e., 1×1, 3×3, 5×5, to cope with multi-scale text instances, which successfully capture characteristics of receptive fields for text. In practice, we factorize the original 5×5 convolutional layer into two convolutional layers with kernel 1×5 and 5×1, which can not only maintain the receptive field but also reduce the number of parameters. Compared with the original RFB, we eliminate a residual shortcut from input to out due to its repeated and same function as residual shortcut existed in structure of TCAM. All those advantages of RFB guarantee that discriminative features can be extracted by this module in an effective way.

Text-Context-Aware Attention Block Considering that a convolutional layer consisting of different channel filters scans the input image and outputs a feature map, each 2D slice of the output 3D feature map essentially encodes the spatial visual responses raised by a filter channel, where the filter performs as a pattern detector, i.e., lower-layer filters detect low-level visual cues like edges and corners, while higher-layer ones detect high-level semantic patterns like parts and objects [17]. By stacking the layers, CNN extracts image features through a hierarchical representation of visual abstractions. Therefore, CNN image features are essentially channel-wise and multi-layer. However, not all the features are equally important and informative for text detection. We thus utilize channel attention scheme to compute task-specified feature map for text detection by exploiting cross-channel relationship. In other words, channel attention scheme aims to involve text context information, i.e., text-specified representation in channel-wise and multi-layer feature map, resulting in a more discriminative feature map of detection task and less unpleasant impact of noise information. We also design a spatial-wise attention scheme by exploring inter-spatial relationship of each channel of feature map, which helps focus on saliency parts of feature map and could act as a beneficial complementary to channel attention.

Based on the above discussions, we thus design TAB as in Fig. 4 (a). Different from former multi-attention method like CBAM [14], we prefer a dual form of the channel and spatial attention scheme, rather than a sequential form of these two schemes. The main reason to apply dual form relies on the purpose to deal with complex and diverse scenes. If sequential contextual embedding is explored, the context from dominated salient objects (e.g., car, building in street scene) will not assist labeling texts, which usually are un-salient in scene. By contrast, in our dual attention model, we selectively aggregate similar semantic features of texts to enhance their feature representations and avoid the influence of salient objects. We thus give the formula of TAB as:

$$A(R(F_l)) = A_c(R(F_l)) \otimes R(F_l) + A_s(R(F_l)) \otimes R(F_l) \quad (3)$$

where function $A_c()$ and $A_s()$ represent process of channel and spatial attention scheme, respectively. To calculate the weight of channel attention, the global average pooling and global max pooling operations are preformed on RFB enhanced feature map $\tilde{F_l}$, which are represented as $F_{a,l}{}^c = AvgP(R(F_l))$ and $F_{m,l}{}^c = MaxP(R(F_l))$. The produced $F_{a,l}{}^c$ and $F_{m,l}{}^c$ have the same size $C\times1\times1$, so they can be seen as two C-d vectors.

As shown in Fig. 4(b), we calculate channel attention weight by firstly concatenating $F_{a,l}{}^c$ and $F_{m,l}{}^c$ as a $2C$-d feature vector. Then, the concatenated feature vector will be fed into a multi-layer percetron (MLP) with two hidden layers. It is noted that the first hidden layer is used to perform a dimension reduction for a compact feature representation to aggregate information of channels. Finally, a sigmoid activation function is used to squeeze the output of MLP. The channel attention weight is represented as:

$$A_c(R(F_l)) = sig(W_1 * (relu(W_0 * [F_{a,l}{}^c, F_{m,l}{}^c]_c))) \quad (4)$$

where $[,]_c$ denotes concatenation operation, function $sig()$ and $relu()$ refer to sigmoid activation function and relu activation function, respectively, W_0 and W_1 are learnable parameter matrices and defined with size $\frac{2C}{r}\times 2C$ and $C\times\frac{2C}{r}$ respectively, and r is a pre-defined dimension reduction parameter and we set it as 16 by experiments.

As shown in Fig. 4(c), the proposed spatial attention firstly performs average pooling and max pooling along channel axis on RFB enhanced feature map $\tilde{F_l}$, which produces $F_{a,l}{}^s$ and $F^s_{m,l}$. The above operations can be represented as $F_{a,l}{}^s = AvgP_c(R(F_l))$ and $F_{m,l}{}^s = MaxP_c(R(F_l))$. Then concatenation operation along channel axis is performed on $F_{a,l}{}^s$ and $F_{m,l}{}^s$. Afterwards, a convolutional layer with 3×3 kernel and a sigmoid function are performed on the concatenated feature map. The spatial attention weight thus can be represented as:

$$A_s(R(F_l)) = sig(Conv([{F_{a,l}}^s, {F_{m,l}}^s]_a) \tag{5}$$

where $[,]_a$ denotes concatenation operation along channel axis, and function $Conv()$ represents the utilized convolutional layer required training.

IV. Experiment

A. Datasets

We evaluated our approach on two widely used datasets, i.e., ICDAR2015 and ICDAR2017-MLT, where the former focuses on incidental text detection, and the latter is used to evaluate performance on multi-oriented and multi-language scene text detection. Specifically, ICDAR2015 contains 1000 and 500 scene images labeled with word-level quadrangles for training and testing, respectively. Meanwhile, ICDAR2017-MLT consists of 7200, 9000 and 1800 scene images for training, testing and validation, respectively.

B. Experiment results and analysis

Table I and II give the detailed statistics of results on ICDAR2015 and ICDAR2017-MLT datasets. It's noted that since ICDAR2017-MLT is a new dataset, we can only find fewer comparative methods. From both tables, we can see that F-measure of the proposed method is the best on both ICDAR2017-MLT and ICDAR2015, which shows the effectiveness of the proposed method. The high F-measure on ICDAR2015 has demonstrated its effectiveness to deal with mulit-oriented texts. The superior performance on ICDAR2017 further verifies its effectiveness for multi-language texts, which further demonstrates the robustness and generalization ability of this method.

The second best precision 0.879 for ICDAR2015 and the second best precision 0.739 for ICDAR2017-MLT prove the significant discrimination power of the feature map in the proposed method. In fact, encoding text context information into feature map by the proposed TCAM mainly improves detection results by eliminating false positive samples, which is shown in Fig. 5. This phenomenon can also be proved by comparing the proposed method with Mask R-CNN, which can simply be regarded as a form of the proposed method without TCAM. Based on the comparisons between these two methods, we can observe the improvement mainly occurs on precision rather than recall values as well. The proposed method achieves the second best and the best recalls on these two datasets, proving the effectiveness of TCAM in distinguishing text and non-text regions.

Sample qualitative detection results are shown in Fig. 5, where the first and second row represent detection results of directly applying Mask R-CNN and the proposed method, respectively. From these sample results, we can see the proposed method detects multi-oriented and multi-language texts well even facing challenges of diversity and complexity in the layout of scene images. Moreover, utilizing TCAM to enhance the representation ability of feature map can highly eliminate false positive samples, thus leading to better detection results.

Table I
Comparison of Performance on ICDAR2015

Method	Precision	Recall	F-measure
FTSN[10]	**0.886**	0.800	0.841
TextBoxes++[1]	0.878	0.785	0.829
PixelLink[18]	0.855	**0.820**	0.837
R2CNN[19]	0.856	0.797	0.825
DDR[20]	0.820	0.800	0.810
EAST[5]	0.832	0.783	0.807
Mask R-CNN baseline	0.863	0.815	0.838
Proposed	0.879	0.816	**0.846**

Table II
Comparisons of Performance on ICDAR2017-MLT

Method	Precision	Recall	F-measure
TDN SJTU2017[21]	0.642	0.471	0.543
SCUT DLVClab1[21]	**0.802**	0.545	0.649
SARI FDU RRPN v1[22]	0.712	0.555	0.624
Mask R-CNN	0.686	0.637	0.660
Proposed	0.739	0.669	**0.702**

C. Implementation Details

All of these experiments are performed on 4 Titan 1080Ti GPUs. We set a single size in {32, 64, 128, 256, 512} for all the anchors on the same pyramid level. In each level of feature pyramid, our model generates 6 anchors with different aspect ratios, i.e., {0.2, 0.5, 1, 2, 5, 7}, on each spatial position. We train this model by SGD with parameters of initial learning rate 0.005, batch size 4, momentum 0.9 and weight decay 0.0005.

V. Conclusion

In this paper, we propose a text-context-aware CNN architecture for mulit-oriented and mulit-language scene text detection. The proposed TCAM can encode context information to enhance feature representation ability, thus improving accuracy of text detector. In the future, we would design a light version to further improve its running speed.

VI. Acknowledgement

This work is supported by the Natural Science Foundation of China under Grant 61672273, Grant 61832008 and Grant 61702160, the Science Foundation for Distinguished Young Scholars of Jiangsu under Grant BK20160021, the Science Foundation of Jiangsu under Grant BK20170892, Scientific Foundation of State Grid Corporation of China (Research on Icewind Disaster Feature Recognition and Prediction by Fewshot Machine Learning in Transmission Lines), and the open Project of the National Key Lab for Novel Software Technology in NJU under Grant K-FKT2017B05.

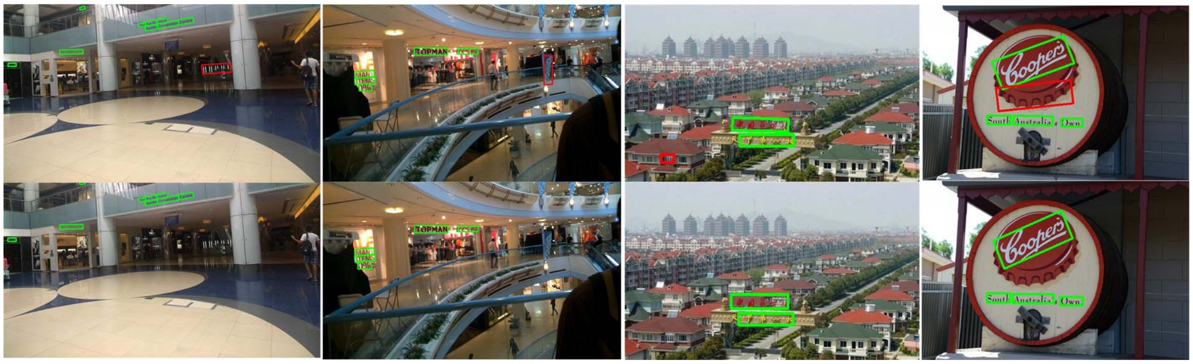

Figure 5. Sample results of directly applying Mask R-CNN (the first row) and the proposed method (the second row). It is noted that the red rectangles refer to false positive detection results by Mask R-CNN, which will be eliminated due to the high distinguish ability of feature map enhanced by TCAM.

References

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

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *Proceedings of ECCV*, 2016, pp. 21–37.

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

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of ICCV*, 2017, pp. 2980–2988.

[5] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "East: an efficient and accurate scene text detector," in *Proceedings of CVPR*, 2017, pp. 2642–2651.

[6] M. Busta, L. Neumann, and J. Matas, "Deep textspotter: An end-to-end trainable scene text localization and recognition framework," in *Proceedings of ICCV*, 2017, pp. 2223–2231.

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

[8] Y. Wu and P. Natarajan, "Self-organized text detection with minimal post-processing via border learning," in *Proceedings of ICCV*, 2017, pp. 5010–5019.

[9] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *Proceedings of CVPR*, 2017, pp. 4438–4446.

[10] Y. Dai, Z. Huang, Y. Gao, Y. Xu, K. Chen, J. Guo, and W. Qiu, "Fused text segmentation networks for multi-oriented scene text detection," in *Proceedings of ICPR*, 2018, pp. 3604–3609.

[11] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao, and G. Li, "Scene text detection with supervised pyramid context network," *arXiv:1811.08605*, 2018.

[12] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in *Proceedings of ECCV*, 2018, pp. 67–83.

[13] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," *arXiv:1611.05594*, 2016.

[14] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of ECCV*, 2018, pp. 3–19.

[15] J. Liu, G. Wang, P. Hu, L. Duan, and A. C. Kot, "Global context-aware attention LSTM networks for 3d action recognition," in *Proceedings of CVPR*, 2017, pp. 3671–3680.

[16] S. Liu, D. Huang, and Y. Wang, "Receptive field block net for accurate and fast object detection," in *Proceedings of ECCV*, 2018, pp. 404–419.

[17] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proceedings of ECCV*, 2014, pp. 818–833.

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

[19] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, "R2cnn: rotational region cnn for orientation robust scene text detection," *arXiv:1706.09579*, 2017.

[20] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," in *Proceedings of ICCV*, 2017, pp. 745–753.

[21] N. Nayef, F. Yin, I. Bizid *et al.*, "Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt," in *Proceedings of ICDAR*, 2017, pp. 1454–1459.

[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 Transactions on Multimedia*, 2018.