MS-ROCANET: MULTI-SCALE RESIDUAL ORTHOGONAL-CHANNEL ATTENTION NETWORK FOR SCENE TEXT DETECTION

Jinpeng Liu *Song Wu* *Dehong He* *Guoqiang Xiao**

College of Computer and Information Science, SouthWest University, Chongqing, China

ABSTRACT

Deep neural networks-based scene text detection has obtained increasing attention in recent years. However, the existing scene text detection methods cannot effectively solve the problem of unclear text features. In this paper, a Multi-scale Residual Orthogonal-Channel Attention Network (MS-ROCANet) is proposed to improve the recall and accuracy of scene text detection. Specifically, a Detail-aware FPN is designed to capture more detailed information. Then, a Shared Composite Attention Head (SCAH) consists of a Residual Orthogonal Attention Module (ROAM), and a Residual Channel Attention Block (RCAB) is proposed. It can enhance textual region features at a multi-scale level. Finally, a global context extraction module is proposed to obtain global contextual information after referring to the core idea of Transformer. Extensive experiments demonstrated that our MS-ROCANet achieved competitive results on a variety of baselines for text detection. The codes and models are available at https://github.com/ASentry/MS-ROCANet

Index Terms— Text Detection, Detail-aware, Cross-scale, Attention, Transformer

1. INTRODUCTION

In recent years, scene text recognition has increased much attention, aiming to detect and label text regions in images with complex backgrounds. It has been widely used in video indexing, machine translation, and video scene parsing. Due to the high sematic information abstraction capability of deep neural networks, the task of scene text detection has obtained higher performance than traditional methods. However, due to the complexity of realistic scenes, how to achieve an effective and efficient scene text detection with complexity situation is still a big challenge.

The existing of recent scenes text detection methods can be roughly divided into two categories: regression-based methods and segmentation-based methods [1, 2, 3, 4, 5]. The regression-based methods design strategies to model the text regions to improve detection accuracy, such as the representative FCENet[6]. However, we have observed that in this kind of research, we only focus on exploring modeling methods and neglect the extraction of more accurate features of text regions. Text region modeling is based on the accurate extraction of features from text regions in images. Still, the complexity of scene text images makes it more challenging to obtain accurate features. Examples of our spotting results are shown in Fig.1. In scene text detection, the k×k convolution kernel of CNN is used to model the texture of text images in any direction. The influence of lighting and blurring causes the neural network to be sensitive to some regions with similar textures as the text regions, resulting in false detection. Another problem is that FCENet only selects the upper three layers of ResNet features as the basis for subsequent modeling. However, we believe that lower-level features will contain richer detail information, crucial for subsequent modeling of text regions.

(a) FCENet (b) ours

Fig. 1. Neural networks that focus on modeling methods alone are prone to the problems of missed detection and false detection in complex scenarios.

In this paper, we propose MS-ROCANet for more accurate scene text detection. The Detail-aware FPN is designed to fuse low-level geometric features with high-level semantic features to obtain a feature map containing richer information. Then, text region features are enhanced at different scales using a Shared Composite Attention Head, and contextual information is obtained by a Global Context Module (GCM). The effectiveness and robustness of the network are confirmed by extensive experiments, which are discussed in Sec.3

Our main contributions in this paper are as follows. 1)

* Corresponding author.

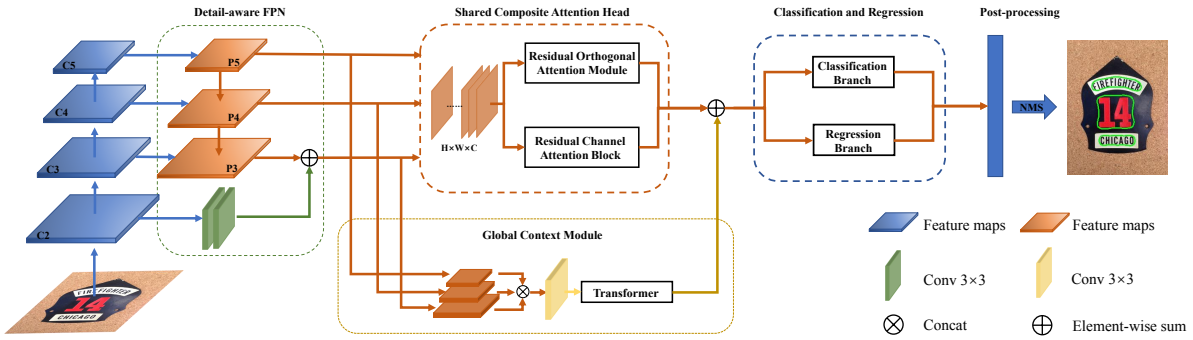

Fig. 2. Overall architecture of our proposed method. Given an image, the feature maps {C2, C3, C4, C5} are extracted by a ResNet50 and a detail-aware FPN. Then, these feature maps are sent into the shared composite attention head to enhance the learning of the text region features. At the same time, the global information is obtained through the global context module. The outputs of the shared composite attention head and the global context module are fused and sent to the classification and regression branches. Finally the results are obtained by post-processing and NMS.

We design a FPN capable of fusing low-level geometric features and high-level semantic features, which is able to obtain more informative feature representation. 2) We propose a novel composite attention mechanism that is more efficient and less computationally intensive. 3) Without relying on external datasets, the proposed method achieves 85.7% and 86.3% of F-measure on CTW1500 and ICDAR 2015 over other methods.

2. PROPOSED METHOD

2.1. Overview

The framework of our proposed method is illustrated in Fig.2. Firstly, the ResNet50 pretrained on ImageNet is leveraged to extract the features of an input image. Afterwards, we construct a FPN-like network to obtain multi-scale features and fuse low-level geometric features with high-level semantic features.Secondly, we design a new shared head, which contains two parallel attention modules, i. e., Residual Orthogonal Attention Module(ROAM) and Residual Channel Attention Block(RCAB), to provide enhanced shared cross-scale features for subsequent classification and regression. Meanwhile, we build a global contextual information extraction module consisting of a Transformer to provide more detailed information for subsequent processing.

2.2. Detailed Architecture

Backbone and Detail-aware FPN. ResNet50 and detail-aware FPN are included in our basic framework. For a given image, we first extract four feature maps $\{C2, C3, C4, C5\}$, which contain {256, 512, 1024, 2048} channels, from different layers of ResNet50, respectively. Unlike the traditional FPN [7], we select the features of the bottom layer of ResNet to obtain geometric features, the upper three layers to obtain semantic features, and fuse the obtained geometric and semantic features to obtain more informative features. The constant convolutional upsampling and downsampling of features during FPN construction will inevitably lead to loss of detailed information, and some of the smaller text regions may even disappear from the final feature map. As the output of lower layer, $C2$ contains richer detail information and more comprehensive target features, so we construct a Detail-aware FPN Module (**DAM**) with two convolutional layers (the kernel size is 3×3 for each larger) to supplement the lost information by FPN, which is shown as follows:

$$P3 = Down_{\times 2}(Conv_{3\times 3}(Conv_{3\times 3}(C2))) + P3 \quad (1)$$

where $Down_{\times 2}$ denotes 2 times down-sampling, $Conv_{3\times 3}$ represents 3×3 convolution.

Shared Composite Attention Head. Due to the complexity of the scene text image itself, the multiple operations of the FPN layer make the features of the text regions more blurred. Therefore, it is necessary to enhance the features of text regions and reduce the influence of non-text regions. Thus, we design a shared composite attention head module (SCAH) to learn robust features for text regions.

As shown in Fig.4, **SCAH** contains two branches: Residual Orthogonal Attention Module (**ROAM**) and Residual Channel Attention Block (**RCAB**). In ROAM, we use asymmetric convolution to model the text regions horizontally and vertically with the same efficiency and lower computational cost. We fuse the obtained feature maps of horizontal and vertical parts as attention maps to weight the text regions, and the formula is shown in eq.2. To guide this learning process, we employ a supervised mechanism to highlight the specified targets. In another branch, we introduce the channel-wise attention (RCAB) to select different semantic features that form a composite attention module with ROAM.

$$F_{ROAM} = (F_{hor} + F_{ver}) \times F_{in} + F_{in} \quad (2)$$

where F_{ROAM}, F_{hor}, F_{ver} and F_{in} denote the ROAM output

(a) CTW1500

(b) ICDAR2015

(c) Total-Text

Fig. 3. Experimental results of our method. (a) results on CTW1500; (b) results on ICDAR2015; (c) results on Total-Text

feature map, horizontal feature map, vertical feature map and input feature map, respectively.

Global Context Module. Benefitting from the ability of buidling long-range dependencies, Transformer has made great progress in the field of computer vision. We believe that building a global information acquisition module using Transformer can effectively improve the quality of text detection. Therefore, by using one of the stages in CrossFormer [8], we build a global context module to acquire the strong features of context in each image. The multi-scale feature maps $P3$ and $P5$ are resized to the same size as $P4$ and then we cascade $P3$, $P4$ and $P5$. Afterwards, the cascaded feature map is fed into the Transformer to obtain global contextual information.

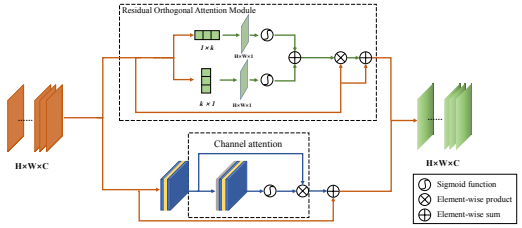

Fig. 4. The visualization of shared composite attention head. The upper branch is ROAM, the lower branch is RCAB.

2.3. Classification and Regression

The classification branch predicts text regions and text center regions, multiplies them by pixels, and obtains the classification score map. The regression branch predicts the Fourier feature vector, which is used to reconstruct the text contour by inverse Fourier transform. Based on the reconstructed contours and classification scores, the final detected texts are obtained by non-maximal suppression (NMS).

Losses. The total loss function of our network can be formulated as follows:

$$L = L_{cls} + \lambda_1 L_{reg} + \lambda_2 hor + \lambda_3 L_{ver} \tag{3}$$

where $L_{cls}, L_{reg}, L_{hor}\ and\ L_{ver}$represent the loss of classification branch, regression branch, horizontal module and vertical module, respectively. L_{cls} and L_{reg} are the same as the classification and regression losses in FCENet. λ_1, λ_2 and λ_3 are three hyper-parameters to balance the four loss functions.

Loss of Horizontal module and Vertical module. L_{hor} and L_{ver} indicate the attention loss to guide the generation of horizontal and vertical attention maps.

$$L_{hor} = \alpha L^b_{hor} + \beta L^d_{hor} \tag{4}$$

$$L_{ver} = \alpha L^b_{ver} + \beta L^d_{ver} \tag{5}$$

where α and β, respectively, indicate the hyper-parametersof the BCE loss $L^b_{hor}, L^b_{ver}$and the dice loss L^d_{hor}, L^d_{ver}.

3. EXPERIMENT

In this section, we first briefly introduce all the datasets used in our experiments. Secondly, the implementation details of our proposed method are presented. Then, we conduct the ablation study and compare our proposed method with several state-of-the-art baseline methods on three datasets.

3.1. Datasets

In our experiments, we utilize three well-known benchmark datasets, i. e., CTW1500 [17], ICDAR2015 [18] and Total-Text [19].

CTW1500 contains both English and Chinese texts with text-line level annotations, where 1000 images for training, and 500 images for testing.

ICDAR2015 is a multi-orientated and street-viewed dataset which consists of 1000 training images and 500 testing images. The annotations are word-level with four vertices.

Total-Text is collected form various scenes, including text-like background clutter and low-contrast texts, with

Table 1. Comparison with related methods on CTW1500, ICDAR2015 and Total-Text, where 'Ext.' denotes extra training data

Methods	Paper	Ext.	CTW1500			ICDAR2015			Total-Text		
			R(%)	P(%)	F(%)	R(%)	P(%)	F(%)	R(%)	P(%)	F(%)
TextSnake [1]	ECCV'18	√	**85.3**	67.9	75.6	80.4	84.9	82.6	74.5	82.7	78.4
SegLink++ [2]	PR'19	√	79.8	82.8	81.3	80.3	83.7	82.0	80.9	82.1	81.5
SAEmbed [3]	CVPR'19	√	77.8	82.7	80.1	85.0	88.3	86.6	-	-	-
CRAFT [9]	CVPR'19	√	81.1	86.0	83.5	84.3	89.8	86.9	79.9	87.6	83.6
PAN [4]	ICCV'19	×	77.7	84.6	81.0	77.8	82.9	80.3	79.4	88.0	83.5
PAN [4]	ICCV'19	√	81.2	86.4	83.7	81.9	84.0	82.9	81.0	**89.3**	85.0
PSENET [5]	CVPR'19	×	75.6	80.6	78.0	79.7	81.5	80.6	75.1	81.8	78.3
PSENET [5]	CVPR'19	√	79.7	84.8	82.2	84.5	86.9	85.7	84.0	78.0	80.9
LOMO [10]	CVPR'19	√	76.5	85.7	80.8	83.5	91.3	87.2	79.3	87.6	83.3
DB [11]	AAAI'20	√	80.2	86.9	83.4	83.2	**91.8**	**87.3**	82.5	87.1	84.7
Boundary [12]	AAAI'20	√	-	-	-	88.1	82.2	85.0	83.5	85.2	84.3
DRRG [13]	CVPR'20	√	83.0	85.9	84.5	84.7	88.5	86.6	84.9	86.5	85.7
ContourNet [14]	CVPR'20	×	84.1	83.7	83.9	**86.1**	87.6	86.9	83.9	86.9	85.4
TextRay [15]	MM'20	√	80.4	82.8	81.6	-	-	-	77.9	83.5	80.6
ABCNet [16]	CVPR'20	√	78.5	84.4	81.4	-	-	-	81.3	87.9	84.5
FCENet [6]	CVPR'21	×	82.8	87.5	85.1	82.6	90.1	86.2	82.7	85.1	83.9
MS-ROCANet	ours	×	83.4	**88.2**	**85.7**	83.2	89.8	86.4	83.3	85.6	84.5

word-level polygon annotations, where 1255 images for training and 300 images for testing.

3.2. Implementation Details

In our method, ResNet50 is used as the backbone with FPN and DCN [20], as shown in Fig.2. During the training stage, we resize the image to 800×800, and data augmentation strategies, including random crop, random rotations, random horizontal flipping, color jitter and contrast jitter are adopted. All experiments are executed on an Nvidia RTX 3060 GPU server. For our proposed method, the batch size is set to 6. Stochastic gradient descent(SGD) is adopted as optimizer with the weight decay of 0.001, and the momentum of 0.9. The initialized learning rate is 0.001, which is reduced 0.8× every 200 epoches. During testing, we resize all the images of three datasets to 1280×800. In this experiment, we set $\lambda_1, \lambda_2, \lambda_3, \alpha$ and β to 1, 1, 1, 0.1, 1 respectively.

3.3. Ablation Study

The purpose of the ablation experiment is to verify the validity of each module, and we demonstrate the validity of each module by comparing it with FCENet on the CTW1500 dataset.

The results of the ablation experiment are shown in Tab. 2. SCAH significantly improved the Recall(R) of CTW1500 by 1.0% over the baseline. In addition, F-meansure(F) on the CTV1500 dataset improved by 0.6%. These performance improvements show that SCAH can effectively improve the recognition of text area features and reduce false detections and missed detections.However, the change in precision was no more than apparent with only a 0.2% increase. Therefore, we introduced two modules, DAM and GCM, and found experimentally that the Precision(P) increased by 0.5%, while the recall remained essentially unchanged. It can be verified that our reinforcement strategy is effective.

Table 2. Ablation study on the three proposed strategies.

SCAH	DAM	GCM	CTW1500		
			R(%)	P(%)	F(%)
-	-	-	82.8	87.5	85.1
√	-	-	**83.8**	87.7	**85.7**
√	√	-	82.4	**89.0**	85.6
√	√	√	83.4	88.2	**85.7**

3.4. Comparisons with State-of-the-Arts

We also compare our methods with previous state-of-the-art methods on several benchmarks. The results are shown in Tab.1. Our method achieves state-of-the-art performance on the datasets CTW1500, Total-Text and competitive results on ICDAR2015. The visualization of multioriented text detection results are shown in Fig.3

4. CONCLUSION

In order to solve the problem of inadequate extraction of text features and lack of detailed information for complex scenes in scene text detection, we propose a novel multi-scale residual orthogonal-channel attention network for scene text detection in this paper. It contains a detail-aware FPN module, a shared composite attention head module, and a global context module. The three modules are integrated for scene text detection with complex backgrounds. Extensive experiments on three benchmark datasets show that our proposed method can effectively enhance the performance and outperforms several state-of-the-art baseline methods in the field of scene text detection.

5. REFERENCES

[1] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao, "Textsnake: A flexible representation for detecting text of arbitrary shapes," in *ECCV*, 2018, pp. 20–36.

[2] Jun Tang, Zhibo Yang, Yongpan Wang, Qi Zheng, Yongchao Xu, and Xiang Bai, "Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping," *Pattern Recognit.*, vol. 96, 2019.

[3] Zhuotao Tian, Michelle Shu, Pengyuan Lyu, Ruiyu Li, Chao Zhou, Xiaoyong Shen, and Jiaya Jia, "Learning shape-aware embedding for scene text detection," in *CVPR*. 2019, pp. 4234–4243, Computer Vision Foundation / IEEE.

[4] Wenhai Wang, Enze Xie, Xiaoge Song, Yuhang Zang, Wenjia Wang, Tong Lu, Gang Yu, and Chunhua Shen, "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *ICCV*. 2019, pp. 8439–8448, IEEE.

[5] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao, "Shape robust text detection with progressive scale expansion network," in *CVPR*. 2019, pp. 9336–9345, Computer Vision Foundation / IEEE.

[6] Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, and Wayne Zhang, "Fourier contour embedding for arbitrary-shaped text detection," in *CVPR*. 2021, pp. 3123–3131, Computer Vision Foundation / IEEE.

[7] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie, "Feature pyramid networks for object detection," in *CVPR*. 2017, pp. 936–944, IEEE Computer Society.

[8] Wenxiao Wang, Lu Yao, Long Chen, Deng Cai, Xiaofei He, and Wei Liu, "Crossformer: A versatile vision transformer based on cross-scale attention," *CoRR*, vol. abs/2108.00154, 2021.

[9] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee, "Character region awareness for text detection," in *CVPR*. 2019, pp. 9365–9374, Computer Vision Foundation / IEEE.

[10] Chengquan Zhang, Borong Liang, Zuming Huang, Mengyi En, Junyu Han, Errui Ding, and Xinghao Ding, "Look more than once: An accurate detector for text of arbitrary shapes," in *CVPR*. 2019, pp. 10552–10561, Computer Vision Foundation / IEEE.

[11] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai, "Real-time scene text detection with differentiable binarization," in *AAAI*. 2020, pp. 11474–11481, AAAI Press.

[12] Hao Wang, Pu Lu, Hui Zhang, Mingkun Yang, Xiang Bai, Yongchao Xu, Mengchao He, Yongpan Wang, and Wenyu Liu, "All you need is boundary: Toward arbitrary-shaped text spotting," in *AAAI*. 2020, pp. 12160–12167, AAAI Press.

[13] Shi-Xue Zhang, Xiaobin Zhu, Jie-Bo Hou, Chang Liu, Chun Yang, Hongfa Wang, and Xu-Cheng Yin, "Deep relational reasoning graph network for arbitrary shape text detection," in *CVPR*. 2020, pp. 9696–9705, Computer Vision Foundation / IEEE.

[14] Yuxin Wang, Hongtao Xie, Zheng-Jun Zha, Mengting Xing, Zilong Fu, and Yongdong Zhang, "Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection," in *CVPR*. 2020, pp. 11750–11759, Computer Vision Foundation / IEEE.

[15] Fangfang Wang, Yifeng Chen, Fei Wu, and Xi Li, "Textray: Contour-based geometric modeling for arbitrary-shaped scene text detection," in *ACM Multimedia*. 2020, pp. 111–119, ACM.

[16] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang, "Abcnet: Real-time scene text spotting with adaptive bezier-curve network," in *CVPR*. 2020, pp. 9806–9815, Computer Vision Foundation / IEEE.

[17] Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Sheng Zhang, "Curved scene text detection via transverse and longitudinal sequence connection," *Pattern Recognit.*, vol. 90, pp. 337–345, 2019.

[18] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman K. Ghosh, Andrew D. Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny, "ICDAR 2015 competition on robust reading," in *ICDAR*. 2015, pp. 1156–1160, IEEE Computer Society.

[19] Chee Kheng Chng and Chee Seng Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in *ICDAR*. 2017, pp. 935–942, IEEE.

[20] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai, "Deformable convnets V2: more deformable, better results," in *CVPR*. 2019, pp. 9308–9316, Computer Vision Foundation / IEEE.