Region-aware Arbitrary-shaped Text Detection with Progressive Fusion

Qitong Wang*, Bin Fu, Ming Li, Junjun He, Xi Peng, Yu Qiao

***Abstract*—Segmentation-based text detectors are flexible to capture arbitrary-shaped text regions. Due to large geometry variance, it is necessary to construct effective and robust representations to identify text regions with various shapes and scales. In this paper, we focus on designing effective multi-scale contextual features for locating text instances. Specially, we develop a Region Context Module (RCM) to summarize the semantic response and adaptively extract text-region-aware information in a limited local area. To construct complementary multi-scale contextual representations, multiple RCM branches with different scales are employed and integrated via Progressive Fusion Module (PFM). Our proposed RCM and PFM serve as the plug-and-play modules which can be incorporated into existing scene text detection platforms to further boost detection performance. Extensive experiments show that our methods achieve state-of-the-art performances on Total-Text, SCUT-CTW1500 and MSRA-TD500 datasets. The code with models will become publicly available at https://github.com/wqtwjt1996/RP-Text.**

***Index Terms*—Scene Text Detection, Scene Understanding, Deep Learning**

I. Introduction

SCENE text detection is a fundamental and challenging task in computer vision community, which focuses on accurately locating text regions in the natural scene. Thanks to the recent development of deep learning technology, scene text detection has witnessed a significant progress and has been widely used in various real-world applications, such as autonomous driving and scene parsing. In recent years, several segmentation-based methods have been developed to generate segmentation masks for text regions, which enable text detection methods to locate arbitrary-shaped text instances. Currently state-of-the-art scene text detection models usually employ FPN [1] originated from general object detection task to extract semantic features and directly concatenate multi-scale information from middle layers, which is a sub-optimal approach for more challenging scene text detection task due to two significant differences between scene text detection and general object detection tasks.

Fig. 1. The challenges in scene text detection task. (a): The complex background information is dominated in natural images. (b): The large scale variance between small and large text instances.

As shown in Fig. 1(a), the complex background information is dominated in scene text detection task and text regions occupy only a small fraction of scene images. Thus, text-related contextual information should be discovered and enhanced in current feature maps. Moreover, different from the objects in general detection task, text instances have large scale variance and the extremely-sized text is common in scene text detection task. For example, in Fig. 1(b), on one hand, the localization information of the small text will be missing in high-level features due to the large down-sampling rate in deep networks. On the other hand, with the limited respective field, the deep network cannot provide enough discrimination information for extremely large text. Therefore, to accurately separate text pixels from natural images, text-related contextual information and multi-scale fusion play important rules.

Based on above discussions, in this paper, we focus on constructing effective multi-scale contextual features for accurately locating text instances in natural images. Specially, to enhance discriminative text-related information for features, we design a Region Context Module (RCM) to adaptively extract region-aware contextual features based on the semantic response in the predefined local regions. In order to gradually recover the missing text information, we further employ multiple RCM branches and propose Progressive Fusion Module (PFM) to fuse semantic contextual information with different scales via a progressive structure. Finally, we incorporate our proposed modules into existing scene text detection platform, termed as "RP-Text", and conduct extensive experiments on four standard text detection benchmarks. Experimental results demonstrate the effectiveness of our proposed methods and we achieve state-of-the-art performances on Total-Text [2], SCUT-CTW1500 [3] and MSRA-TD500 [4] datasets.

Manuscript received March 25, 2021; accepted May 22, 2022. This work is partially supported by the Joint Lab of CAS-HK, the Shenzhen Research Program (JSGG20191129141212311, RCJC20200714114557087), the Shanghai Committee of Science and Technology (Grant No. 21DZ1100100).

*Work done during an internship at Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences.

Q. Wang and X. Peng are with the Department of Computer and Information Sciences, University of Delaware. DE, USA. (email: {wqtwjt@udel.edu, xipeng@udel.edu})

B. Fu, M. Li, J. He and Y. Qiao are with Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China. (email: {bin.fu@siat.ac.cn, ming.li3@siat.ac.cn, hejunjun@sjtu.edu.cn, yu.qiao@siat.ac.cn})

Y. Qiao is also with Shanghai AI Laboratory, Shanghai, China.

Q. Wang and B. Fu are equally-contributed authors.

Y. Qiao is the corresponding author.

The contribution of our work can be summarized as follows:

• We introduce a Region Context Module (RCM) to extract and enhance text-related region-aware contextual information, which usually occupies only a small fraction of images, according to the semantic response in the predefined local regions.

• To deal with scale variance problem in text detection, the Progressive Fusion Module (PFM) is further developed to gradually fuse multi-scale contextual information with multiple RCM branches.

• Our proposed modules are the plug-and-play modules. We construct our RCM and PFM on existing text detection platforms to evaluate the effectiveness of our methods. We achieve state-of-the-art performances on Total-Text, SCUT-CTW1500 and MSRA-TD500 datasets.

II. Related Work

A. Scene Text Detection

Current state-of-the-art text detection methods can be roughly divided into two different categories: regression-based approaches [6]–[14] and segmentation-based approaches [5], [15]–[23]. The standard philosophy of regression-based approaches is to formulate text region as rectangular bounding boxes and optimize deep network to regress position coordinates of text instances. TextBoxes++ [11] proposes an end-to-end trainable deep neural network to detect oriented text from scene natural images based on SSD [24]. DMPNet [13] relies on quadrilateral sliding windows and Monte-Carlo method to generate initial text bounding coordinates, then moderately regress coordinates to obtain final prediction results. CTPN [14] employs an Long Short-term Memory (LSTM) [25] to explore rich sequential signals in natural scenes, for effectively predicting horizontally text regions. To match the rotated text regions, RRPN [26] introduces a series of rotated anchors to predict candidate text regions. SPCNet [9] builds on a branch for text region semantic segmentation on the basis of Mask-RCNN [27], and merges the intermediate features of semantic segmentation with the features of the detection branch, then multiplies the prediction results of semantic segmentation as an attention mask back to the feature map. Recently, several regression-based methods have been put forward to modify rectangular bounding boxes to predict arbitrary-shaped regions, such as CounterNet [6], ABCNet [28] and ATTR [8].

Thanks to the pixel-level prediction, segmentation-based framework is another reliable paradigm which also can handle arbitrary-shaped text regions, and thus have rapid development in recent works. EAST [23] proposes a two-stage text detection method, Full Convolutional Network (FCN) [29] based detection and proposed Locality-Aware Non-Maximum Suppression (LA-NMS) algorithm, which eliminates intermediate process redundancy and reduces detection time. PixelLink [22] predicts the connection relationships between neighborhood pixels, which enable this model to accurately separate closed text instances. TextSnake [20] proposes a snake-like text representation to solve the arbitrary-shaped text detection problems in natural scenes. CRAFT [17] performs character-wise text detection by predicting the single-character Gaussian heatmap and the connectivity between characters. For the datasets without character-level annotations, a weakly-supervised method is further employed to generate character-level pseudo ground truth. PSENet [18] predicts kernels with different shapes for text regions and expands from minimal kernel to form final text regions. PAN [5] further develops PSENet by speeding up the time-consuming feature extraction backbone and post-processing to ensure the real-time characteristic of detecting text in natural scenes. Recently, DRRG [15] detects arbitrary-shaped text regions based on the connection of character and text components, with the help of Graph Convolutional Network (GCN) [30] to reason about deep relationships between different components of text regions, effectively solving the connection problem of text region components in natural scenes. To our best knowledge, no existing work has carefully discussed both the dominant background and substantial scene text variance challenges.

B. Attention and Gating Mechanism

Similar with people paying attention to attractive objects in surroundings, attention mechanism is designed to assign distinctive attention weights to different regions for input images. The attention mechanism has been widely employed in recurrent neural network based methods to extract discriminative features for each time step, such as LSTM [25] and GRU [31]. Recently, attention mechanism has been successful applied in computer vision community. Non-local Neural Network [32] introduces long-range dependencies into features according to the affinity relations. Gated-CNN [33] employs gate mechanism to extract and process boundary-related information. ACNet [34] develops adaptive context blocks to efficiently extract relative local and global signals for scene parsing. However, these methods cannot be directly applied for scene text detection since we need to extract text instances from only a few image pixels. Therefore, a new region-aware based gate mechanism should be carefully explored in scene text detection field.

C. Multi-scale Fusion

Although deep layers in CNN contain rich semantic information, the local information is missing in high level features, which is a serious problem in per-pixel prediction task. Moreover, it becomes more serious in scene text detection task, since some small text regions are lost by the large down-sampling rate. To remedy this problem, most segmentation-based text detection frameworks fuse different scale feature maps step by step [17], [20], [23]. Moreover, PSENet [18] concatenates feature maps with different scales directly to generate final feature maps before detection head module. Therefore, how to optimally integrate feature maps with different scales is still an open problem in scene text detection researches.

Considering the characteristics of the scene text detection, we design a progressive fusion module and a region-based gate module for text detection to efficiently extract multi-scale region-aware features in natural scene images. To the best of our knowledge, our paper is the first study to explore

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3181448

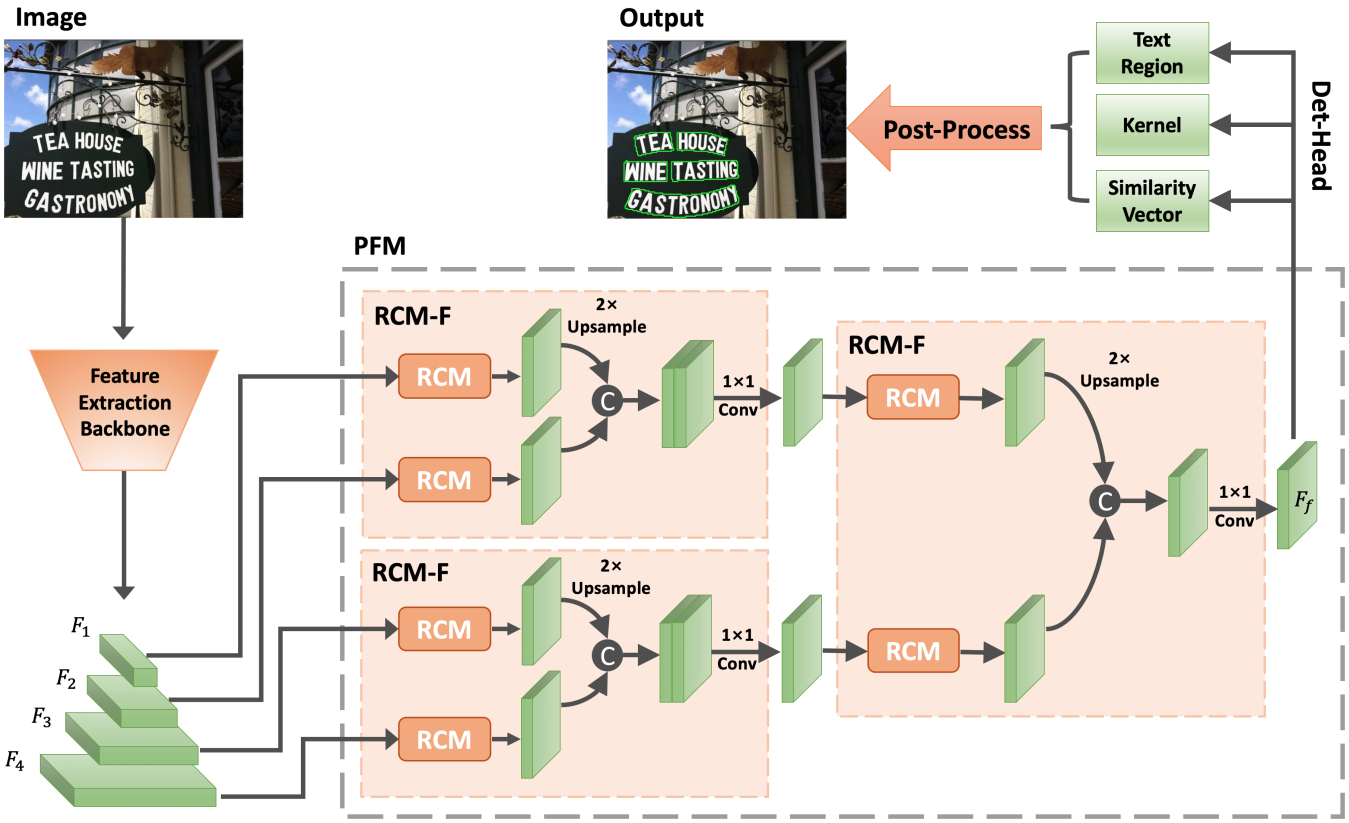


Fig. 2. Overall pipeline of our RP-Text model. We employ PAN [5] as our backbone for feature extraction. The Region Context Module (RCM) is developed to extract and enhance text-related features while the Progressive Fusion Module (PFM) is proposed to collect multi-scale information in an iterative manner. Following PAN [5], the resulted feature F_f is passed to Det-Head for predicting text regions, kernels and similarity vectors, which will be utilized to generate final text regions by post-processing. "RCM-F" denotes two "RCM" in parallel followed by 2 $\times$ upsampling for relatively small feature maps, concatenations and 1 $\times$ 1 convolution operations. "C" denotes concatenation operation.

the effectiveness of self-attention mechanisms to solve the background complexity and scene text scale variance in scene text detection field. The relevant details will be given in following section.

III. Methodology

We aim at solving the challenging problems in scene text detection: background complexity and scene text scale variance. The key idea is to extract scale-specific text-related features in local regions and then employ multiple Region Context Module (RCM) branches to collect multi-scale contextual information via Progressive Fusion Module (PFM). Finally, as shown in Fig. 2, we incorporate our modules into an existing text detection platform and introduce our overall pipeline.

A. Region Context Module

Contextual relations can provide rich discriminative information for per-pixel prediction, which is crucial for separating text/non-text regions from complex background pixels in scene text detection task. For a well-optimized deep network, it will gradually extract abstractive semantic information from visual details, and thus the pixels belong to text regions will have large response value in feature maps. Moreover, unlike general object detection and segmentation, text instances usually locate in limited regions. Since self-attention mechanism can effectively extract essential information from feature maps, we separate one feature map into predefined local regions and adaptively enhance text-related region-aware contextual information. Besides, in most natural scenes, background pixels are dominated while some text regions may locate in inconspicuous positions. Based on the abovementioned observations, a self-attention gate with max-pooling operation is utilized to filter out unrelated background regions and highlight text regions.

Therefore, we separate the feature map into predefined $\mu \times \mu$ local regions, and adaptively enhance text-related region-aware contextual information based on the corresponding response value. In the following, we introduce our Region Context Module (RCM) in details.

As shown in Fig. 3, given an input feature map $I \in R^{H \times W \times C}$, we first separate I into a set of predefined $\mu \times \mu$ local regions $I_\theta \in R^{\frac{H}{\mu} \times \frac{W}{\mu} \times C}$ (we set $\mu = 8$ in this paper). Then a max pooling operation followed with a 1×1 convolution is utilized to obtain region-specific response $F_\theta \in R^{1 \times 1 \times C}$ by extracting and refining highest semantic response for each sub-region. We employ gate mechanism to adaptively extract and enhance text-related features according to similarity matrix between the resulted region-specific response F_θ and the corresponding pixels. For the pixel i in local region θ, the

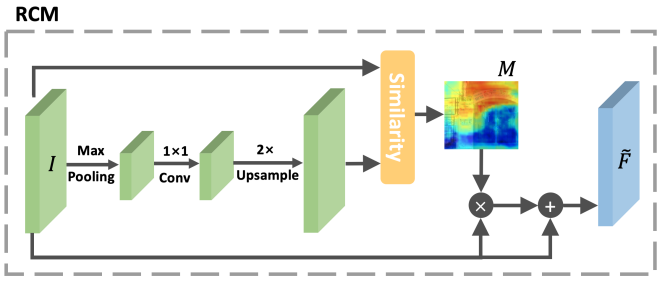


Fig. 3. Architecture of our proposed Region Context Module (RCM). "M" denotes generated similarity matrix for extracting and enhancing text-related information in input features. "$\times$" denotes element-wise multiplication and "$+$" denotes weighted summation operation.

corresponding element of similarity matrix can be formulated as

$$M_i^{\theta} = \exp\left(-\frac{1}{\delta}||I_i - F_{\theta}||_2\right) \tag{1}$$

where we employ L2 norm $||\cdot||_2$ to measure distance between I_i and the region-specific response F_{θ}. The exponential function is utilized to normalize each element in similarity matrix. δ is a hyper-parameter to rescale distance and we take $\delta = 4$ in following experiments.

According to the resulted similarity matrix, the text-related features can be enhanced via gate mechanism and weighted residual connection, which can be expressed as

$$\tilde{F} = \gamma \times M \times I + I \tag{2}$$

where γ is a learnable parameter to balance the contribution between input features I and the enhanced features.

B. Progressive Fusion Module

In previous section, we have developed Region Context Module (RCM) to extract distinguishable text-related contextual information in a set of predefined local regions. However, due to the large scale variance of text instances in natural scene, a single scale-specific RCM may miss some pivotal text-related features for detecting text instance, especially for the small text regions in high-level feature maps. To remedy this problem, in this paper, we employ multiple RCM branches to extract multi-scale contextual information from different layers and gradually combine them via progressive fusion.

As shown in Fig. 2, from feature extraction decoder, we build our RCM branches on top of four feature maps F_1, F_2, F_3, and F_4 to extract multi-scale text-related features. Then, instead of concatenating features of different levels directly, we separate the resulted feature maps into two pairs and merge them in an iterative manner. Specially, we first concatenate F_1 with F_2 and F_3 with F_4 followed with a 1×1 convolution to fuse different features. Then two RCM branches are further constructed on the resulted feature maps F_l and F_h to fuse them in the similar manner.

From our perspective, the advantages are twofold:

• It merges feature maps with different scales respectively. As a result, the features can be better integrated and the contextual information of different scales provides complementary information for accurately detecting text regions.

• With the help of the progressive structure of PFM, the overall structure of our neural network becomes deeper, introducing more nonlinear features, which may generate more robust and discriminative features. Therefore, it enables our framework to improve the performance of scene text detection.

C. Discussions

Comparison with state-of-the-art segmentation-based text detection methodologies: Recent state-of-the-art text detection works pay limited attention to extract contextual features, which can provide rich discriminative information for per-pixel prediction and are crucial for separating text/non-text regions from complex background pixels. In recent state-of-the-art segmentation-based studies [18], [20], [23], people mainly focus on constructing complex text region representations and ignore the importance of specific deep modules for text detection task. In our current work, RCM is proposed to adaptively extract region-aware contextual features based on the semantic response in the predefined local regions. To the best of our knowledge, our paper is the first study to explore the effectiveness of region-aware based gate mechanisms in the field of scene text detection.

Comparison with state-of-the-art multi-scale methodologies: Detection from multi-scale perspective has also studied in both scene text detection and general object detection frameworks. From scene text detection perspective, state-of-the-art methodologies are different from ours even though they extract multi-scale features to improve robustness of their models. In order to obtain features of different scales, MSR [19] takes as input the different scales of the images into the deep network, respectively outputs the predictions for text regions. But this calculation resource consumption is too enormous to be realistic in application. [5], [18] concatenates feature of different levels directly in the last step from deep network perspective. However, we think that concatenating features of different levels directly is not completely potential to detect text regions from multi-scale perspective. From general object detection perspective, Cascade-RCNN [35] is composed of a series of detection modules, each of which is trained based on positive and negative samples with different IoU thresholds. The output of the previous detection module is employed as the input of the latter detection module, where stage-by-stage training strategy is implemented to extract multi-scale positive samples under multiple IoU thresholds. There are two differences between Cascade-RCNN and our current work. Firstly, Cascade-RCNN uses multi-scale structure to iteratively regress proposals while our PFM utilize multi-scale structure to fuse feature maps with different scales. Secondly, Cascade-RCNN iteratively regresses proposals based on a single feature while our PFM fuses different features with multiple scales to reach our text detection goal, and thus our PFM is able to utilize more plentiful semantic features to generate accurate predictions. In our proposed PFM, instead of concatenating features of different levels directly, we separate the resulted feature maps into two pairs and merge them in an iterative manner. Not only we find that this way make it possible that the contextual information of different scales provides

complementary information to each other, but also we think that the progressive structure of our proposed PFM makes neural network deeper and provides more nonlinear features, which provide more robust and discriminative predictions.

IV. Experiments

In this section, we perform extensive experiments to demonstrate the effectiveness of our method on four standard scene text detection benchmarks, including Total-Text [2], SCUT-CTW1500 [3], ICDAR2015 [36] and MSRA-TD500 [4]. Experimental results demonstrate that our proposed model achieves state-of-the-art performance on Total-Text, SCUT-CTW1500, MSRA-TD500 datasets. In the following, we will introduce implementation details of our model and then perform several ablation experiments on Total-Text dataset. Then, the experimental results on four standard benchmarks are given. Finally, we provide visualizations results and speed analysis of our method.

A. Datasets

SynthText [37] is a large scale text detection dataset with $800,000$ synthetic images which are created by including approximately 8 million quadrilateral synthetic text instances with random fonts, sizes, colors on background images. It takes into account the natural scene layout.

Total-Text [2] contains $1,255$ training and 300 testing images with polygons annotations. These images have more than 3 different directions of text regions: horizontal, multi-directional and curved. The images are collected from real scenes and contain horizontal, multi-oriented and curved text instances.

SCUT-CTW1500 [3] is another well-known standard arbitrary-shape text detection benchmark which contains $1,000$ training and 500 testing images with English and Chinese scripts for curved text detection. Every text annotation is marked as 14-point polygon.

ICDAR-2015 [36] is one of commonly used datasets in scene text detection domain, which includes $1,000$ training and 500 testing images collected from Google Glass. The text regions are annotated by the four vertices of quadrilateral.

MSRA-TD500 [4] focuses on multilingual quadrilateral text in natural scenes. Text instances of indoor images are mainly from signs, door panels and warning signs, while those of the outdoor images are mainly from brands and billboards under complex natural scenes. The resolution of images is between 1296×864 and 1920×1280. It contains 300 training and 200 testing images with English and Chinese scripts.

HUST-TR400 [38] comprises 400 images which include English scripts and Arabic numbers with different fonts, sizes, colors and orientation. Following PAN [5], to solve the training data insufficiency problem of MSRA-TD500, we take HUST-TR400 as extra training data when fine-tuning on MSRA-TD500.

B. Implementation Details

To evaluate our plug-and-play modules (RCM and PFM), we choose PAN[1] [5] as our baseline model, and incorporate our proposed modules into the baseline model, termed as RP-Text. The overall pipeline is shown in Fig. 2. In original PAN [5] model, the four feature maps after FFM addition module are upsampled and concatenated into a final feature map with 4×128 channels. In this paper, we remove the upsampling and concatenating operation. Instead, we implement our RCM and PFM on top of the "element-wise addition" layer of FFM in PAN [5]. Following PAN [5], the output feature map of our proposed module is further employed to predict text region, kernel and similarity vectors for clustering pixels into text regions. We implement the same loss function as PAN [5] to optimize our model:

$$L = L_{tex} + \alpha * L_{ker} + \beta * (L_{agg} + L_{dis}) \tag{3}$$

where L_{tex} is the dice loss [39] of text regions and L_{ker} is the dice loss of text kernels. L_{agg} denotes aggregation loss which clusters text regions and kernels in the same text instances while L_{dis} denotes discrimination loss to keep the distance among the kernels. α and β are used to balance the importance among elements of loss function.

We apply Adam [40] optimizer in our experiments. For the optimization process, we first pretrain our model on SynthText dataset with one epoch, then fine-tune on specific benchmarks with 600 epochs, such as Total-Text, ICDAR2015 and SCUT-CTW1500. Since the training set of MSRA-TD500 is rather small, we follow PAN [5] to include 400 images from HUST-TR400 [38] as extra training data.

In training process, we utilize 2 GPUs with 8 images per GPU for Total-Text and SCUT-CTW1500 datasets while 4 GPUs with 4 images per GPU for ICDAR-2015 and MSRA-TD500 datasets. In testing stage, single-scale testing strategy is employed. The short sides of testing images are kept as 640 for Total-Text dataset and 736 for MSRA-TD500 dataset, which keep the same with PAN [5]. Besides, the short sides of input images are kept as 700 for SCUT-CTW1500 dataset and 1152 for ICDAR-2015 dataset. The down-sampling backbone, learning rate strategy, kernel settings and negative positive ratio of Online Hard Example Mining (OHEM) [41], are kept the same as PAN [5]. The above implementation details apply to two scenarios: (1) Our RP-Text model. (2) For the sake of complete fairness, we reimplement PAN [5] in standard benchmark datasets under the same conditions.

C. Ablation Analysis

In this section, we conduct extensive experiments on Total-Text dataset [2] to validate our proposed model with different settings and demonstrate the effectiveness of our model. We employ PAN [5] as our baseline and reimplement this model for fair comparison, which is marked as † in TABLE II III IV.

1) Ablation Study for RP-Text: In the following, we evaluate the contribution of RCM and PFM in our RP-Text method.

[1]The codebase is official PyTorch inplementation: https://github.com/whai362/pan_pp.pytorch

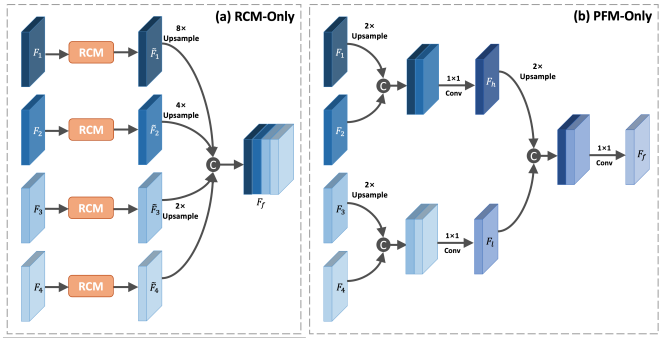


Fig. 4. Architecture of (a): RP-Text w/o PFM; (b): RP-Text w/o RCM, in our ablation study. "C" denotes concatenation operation in this figure.

RP-Text model without RCM and PFM modules is the PAN [5] model which serves as the baseline for comparison in this section. For RP-Text without PFM, as shown in Fig. 4(a), we employ four RCM branches to extract scale-specific text-related context features on F_1, F_2, F_3 and F_4, and then concatenate the resulted feature maps directly, which follows [5], [18] from multi-scale detection perspective. For RP-Text without RCM, as shown in Fig. 4(b), we simply concatenate different feature maps followed with 1×1 convolution in a progressive manner. Finally, we combine RCM and PFM modules with baseline model to construct our RP-Text model.

Experimental results are shown in TABLE I and several conclusions can be drawn from this table:

• When RCM exists, with the help of PFM, there are 1.7% improvement on precision while recall increased by 0.9%, which verify that the multi-scale information has a positive contribution to detecting text regions with different scales. One the one hand, from Fig. 5, we find that RP-Text can better detect both large and small text regions. (red rectangle of row. 1) But RP-Text without PFM only detects text regions from single scale. (red rectangle of row. 2) The small text regions are detected as one big text instance whose predicted size is quite similar with neighborhood predictions, which is called as "one-to-many" detection [42]. In our ablation analysis, the evaluation calculation formula for Total-Text [2] dataset can be defined as

$$M_{i,j} = \frac{S(gt_i \bigcap det_j)}{S(det_j)} \tag{4}$$

where $M_{i,j}$ denotes element of precision calculation matrix that stores the precision rate of each detection boxes, gt_i denotes ith ground-truth bounding text region, det_j denotes jth bounding text region prediction result, S denotes pixel area of text region, $\bigcap$ denotes intersection pixel area of two regions. From this equation, we find that when $S(gt_i \bigcap det_j)$ keeps unchanged, the larger the $S(det_j)$, the lower the value of $M_{i,j}$, which in turn lowers the precision score. On the other hand, since RP-Text can better detect text regions with multiple scales, the performance of detect small text regions can improve as well (see the blue rectangles of row. 1 and row. 2 in Fig. 5) so that recall can be improved. This experiment shows that concatenating features of different levels directly which follows [5], [18] is a sub-optimal option. Our PFM, however, behaves superior when predicting text instances of multiple scales.

TABLE I
EXPERIMENTAL RESULTS WITH DIFFERENT COMBINATIONS OF RP-TEXT. COMBINED WITH RCM AND PFM, OUR RP-TEXT SIGNIFICANTLY IMPROVES TEXT DETECTION PERFORMANCE ON TOTAL-TEXT DATASET.

Baseline	RCM	PFM	P (%)	R (%)	F (%)
√			88.4	80.3	84.2
√	√		87.7	81.9	84.7
√		√	89.0	80.3	84.5
√	√	√	89.4	82.8	86.0

• When PFM exists, with the help of RCM, recall increases by 2.5% while precision only increases by 0.4%, which indicates our RCM module can effectively fill the text areas missed by PAN [5]. From Fig. 5, we find RCM helps RP-Text detect small text instances with better accuracy (see the red rectangles of row. 1 and row. 3). For small text regions, the characteristic that locating in limited regions of images is exceptionally obvious so that RCM's ability that filling the small text instances missed by the baseline verifies that our RCM is expert at extracting local text-related signal from natural scenes with dominating background.

• When combining PFM and RCM together, there is a significant improvement 1.8% on F-measure. Therefore, PFM and RCM can provide complementary context information for scene text detection task. With the help of PFM, RCM can focus on multi-scale regions and adaptively extract text-related features for each scale. With the help of RCM, PFM can effectively detects text in differently scales in more potential predicted text regions.

2) Ablation Study for Region Context Module: In this section, we focus on studying the influence of Region Context Module (RCM) with different structures and settings.

Effective Pooling Method for RCM: In this paper, we employ max pooling operating to distinguish and extract text-related regions in our RCM module. However, in recent works [34], [43]–[45], average pooling is another popular operation to adaptive extract global contextual information, especially for the pixels belong to background or dominated objects. We perform extensive experiments to indicate the differences between max pooling and average pooling. As shown in TABLE II, max pooling outperforms average pooling by 1.2% in F-measure in RCM module. Comparison visualizations between and average pooling and max pooling are shown in Fig. 6.

However, RCM with average pooling still outperforms PAN [5] baseline by 0.6% in F-measure. We think the reason why RCM with average pooling still outperforms PAN [5] baseline is owing to the existence of predefined pooling local regions.

Effectiveness of Predefined Local Regions in Proposed RCM: In order to further verify the effectiveness of our proposed predefined pooling local regions for text detection, we also provide experiment results of global average pooling based RCM, which completely follows [34], [43]–[45]. According to TABLE II, RCM with global average pooling even deteriorate 0.3% points in F-measure comparing with baseline.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3181448

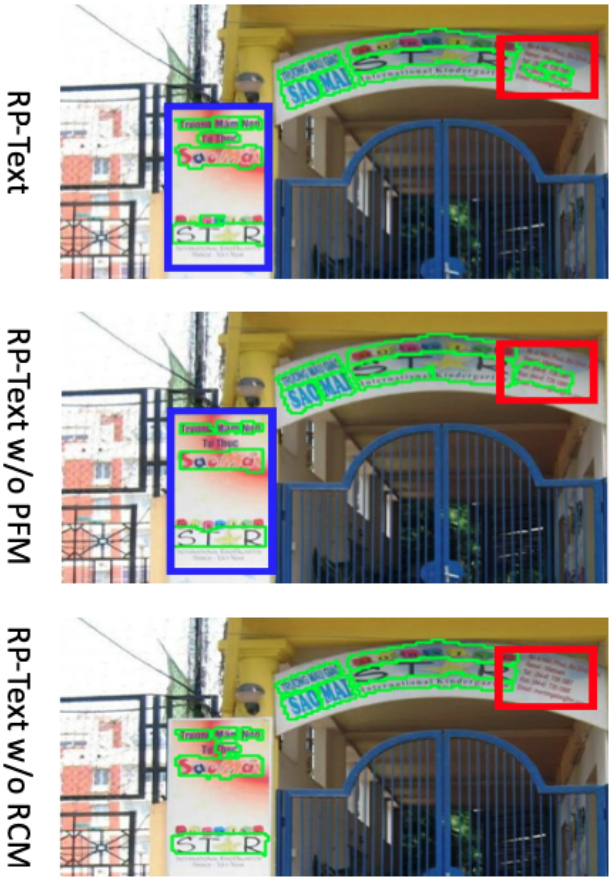


Fig. 5. Prediction visualization for RP-Text (with both Region Context Module (RCM) and Progressive Fusion Module (PFM)), RP-Text without PFM, RP-Text without RCM. The main differences are marked using red and blue rectangles.

TABLE II
COMPARISON BETWEEN DIFFERENT POOLING TYPES IN RCM. GLOBAL AVG DENOTES GLOBAL AVERAGE POOLING LAYER WHICH IS BROADLY USED IN STATE-OF-THE-ART SEGMENTATION METHODS TO EXTRACT GLOBAL FEATURES. AVG AND MAX DENOTE AVERAGE POOLING AND MAX POOLING IN RCM, RESPECTIVELY. IN THIS CASE, WE SET THE POOLING SIZE AS 8 IN EXPERIMENTS.

Methodology	Pooling Type in RCM	P (%)	R (%)	F (%)
PAN [5] †	-	88.4	80.3	84.2
RP-Text	Global Avg	88.3	80.0	83.9
	Avg	87.8	82.1	84.8
	Max	89.4	82.8	86.0

This experiment studies that when negative pixel signal is dominated in natural images, max pooling can be applied to extract local text-related signal from scene text images. Different from semantic segmentation tasks, background information is completely noisy signal in natural scene images for text detection. Applying average pooling cannot give promising improvement since it introduces noise signal from background pixels when applying self-attention mechanism. Besides, different from [34], [44]–[46], predefined local regions are separated to help extract textual information before operating pooling, which enable RCM to extract region-aware signal in limited regions with more efficiency.

Visualization Analysis of similarity matrix in our Region Context Module: From Fig. 6, we find that the generated self-attention gates indeed have higher response to the text instances, which enable our RP-Text to strengthen the weight for text regions when predicts text polygons. Moreover, Fig. 6 also reveals that our Region Context Module with max pooling has stronger signal response in text regions than that with average pooling.

TABLE III
ABLATION STUDY FOR SUITABLE POOLING SIZES IN RCM. THE POOLING TYPE IS MAX POOLING IN EXPERIMENTS.

Methodology	RCM pooling size	P (%)	R (%)	F (%)
PAN [5] †	-	88.4	80.3	84.2
RP-Text	2	87.7	80.6	84.0
	4	89.3	81.4	85.2
	8	89.4	82.8	86.0
	16	88.4	81.4	84.7

TABLE IV
COMPARISON BETWEEN DIFFERENT FUSION STRUCTURES IN PROGRESSIVE FUSION MODULE.

Methodology	PFM type	P (%)	R (%)	F (%)
PAN [5] †	-	88.4	80.3	84.2
RP-Text	A	89.4	82.8	86.0
	B	88.2	81.1	84.5
	C1	88.6	81.1	84.7
	C2	88.7	82.0	85.2

Suitable Pooling Size in Region Context Module: As we have discussed in Sect. III-A, compared with general object detection, text instances usually locate in limited regions and we separate feature maps into $\mu \times \mu$ local regions to better capture context information. Therefore, the hyper-parameter μ is a significant parameter for our model, which provides an important scale priori about text regions. Experiments in TABLE III verify this perspective and our method achieves best performance when pooling size is 8. We will keep this setting in following experiments.

3) Ablation Study for Progressive Fusion Module: In this section, we focus on studying our Progressive Fusion Module (PFM) with four different structures for multi-scale feature fusion, termed as PFM-A, PFM-B, PFM-C1 and PFM-C2 (see Fig. 7 for details), respectively. PFM-A is our default choice in other experiments. It is worthy to mention that the PFM-C1 and PFM-C2 have the same structure with different fusion order.

Experiment results are present in TABLE IV and the following conclusions can be drawn: (1). Compared with PFM-B, the improvement of PFM-A (1.8%) is more considerable. (2). For PFM-B, the improvement is very limited comparing with PAN [5] baseline in F-measure (0.3%). We think the reason may come from the information redundancy and the training progress complexity in PFM-B. (3). Compared with PFM-C1, PFM-C2 achieves better performance, since the finest feature map (F_4) can provide more local text-related response and provide better guidance for further fusion. Based on above experiments, we select PFM-A as our final Progressive Fusion Module and employ this structure in the following experiments.

Visualization Analysis of PFM: We visualize attention

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3181448

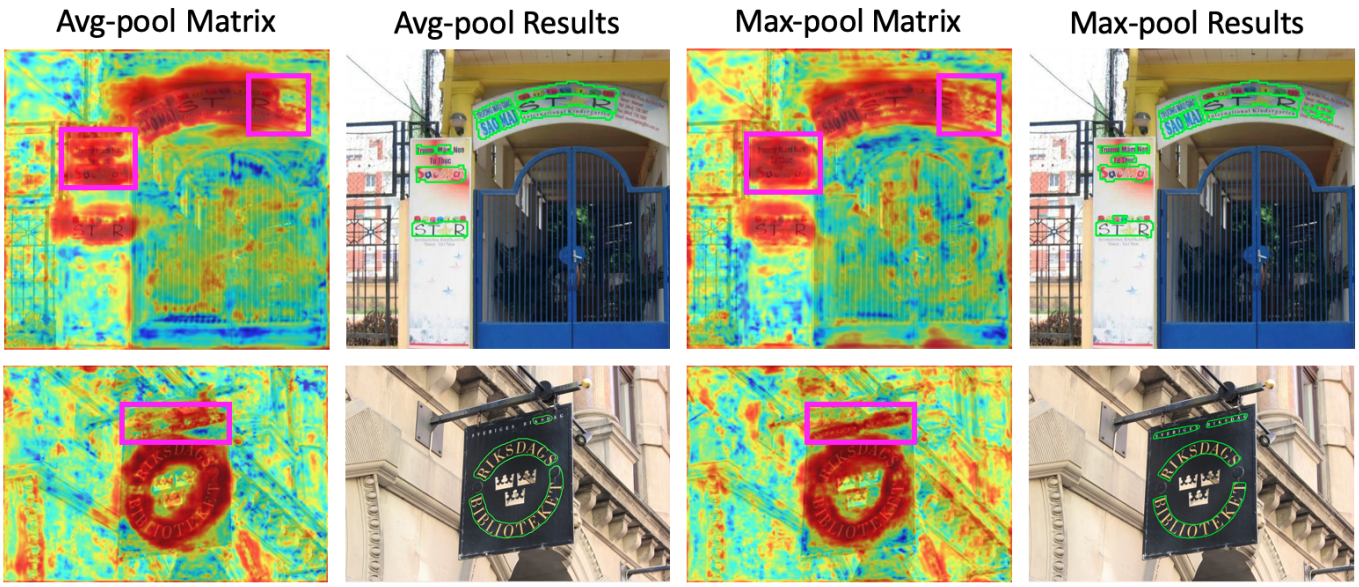


Fig. 6. Visualization results of similarity matrix in our Region Context Module (RCM). The first and third columns denote generated self-attention gates and the second and forth columns denote original natural scene images with predicted bounding polygons. Green polygons denote bounding polygon predictions. By comparing through similarity matrix with average pooling and that with max pooling, when predefined local regions exist, RCM with max pooling has stronger signal response in text regions than that with average pooling (see pink rectangles) so that the detection results with max pooling are better.

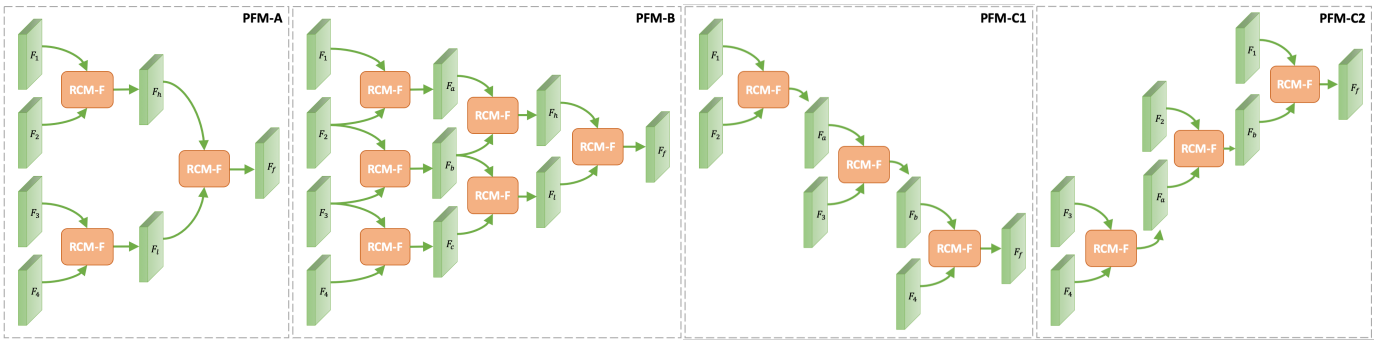


Fig. 7. Different structures of Progressive Fusion Module (PFM) in ablation study. The Four feature maps, F_1, F_2, F_3, F_4, denote outputs from our backbone for feature extraction. F_a, F_b, F_c, F_l, F_h serve as the middle-of-the-road feature maps for the resulted feature F_f. The "Region Context Module Fusion" (RCM-F) denotes two Region Context Module (RCM) in parallel followed by 2 $\times$ upsampling for relatively small feature map, concatenation and 1 $\times$ 1 convolution operation. The detailed structure of RCM-F module is shown in Fig. 2.

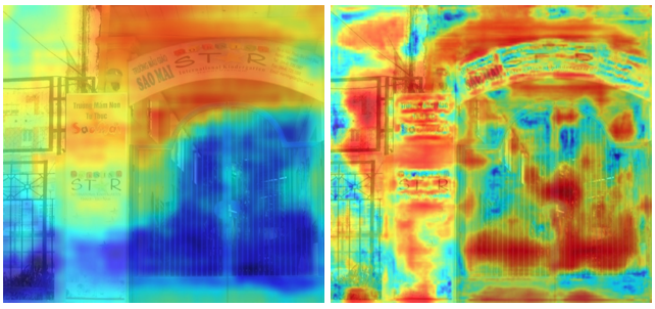


Fig. 8. Comparison of visualization results of similarity matrix in our Region Context Module (RCM). Left: high-level gate after fusion F_1 and F_2 in Fig. 2. Right: low-level gate after fusion F_3 and F_4 in Fig. 2.

map of high-level features (after fusion F_1 and F_2 in Fig. 2) and low-level features (after fusion F_3 and F_4 in Fig. 2). As shown in Fig. 8, the high-level gate can accurately highlight text regions with low resolution while low-level gate enhances detailed text instance signal and contains much unrelated noise. Fortunately, these noise can be depressed by high level gate.

D. Module Capacity Analysis

As we discussed in previous chapters, RCM and PFM help to achieve better performances than baseline framework. However, it is possible that the improvements are due to more parameters of our modules. To further verify the effectiveness of our architectural design, we further study the capacity of our modules by adding more parameters to baseline framework to reach approximately the same parameters as our RCM and PFM.

The total parameter volume of our proposed RCM and PFM is 211K. We reimplement the baseline framework with

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3181448

Fig. 9. Comparison of text detection results between PAN [5] and our RP-Text.

TABLE V
MODULE CAPACITY ANALYSIS FOR OUR RP-TEXT. "PAN" DENOTES ORIGINAL BASELINE, "PAN-M" DENOTES BASELINE FRAMEWORK WITH MORE PARAMETERS WHICH KEEP THE ROUGH SAME PARAMETERS AS OUR PROPOSED RCM AND PFM.

Dataset	Total-Text [2]			SCUT-CTW1500 [3]			ICDAR-2015 [36]			MSRA-TD500 [4]		
Methodology	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
PAN	88.4	80.3	84.2	**88.0**	80.1	83.9	88.3	81.1	84.5	87.5	83.0	85.2
PAN-m	87.6	79.9	83.6	84.7	81.5	83.0	86.3	80.5	83.3	87.7	83.1	85.3
RP-Text (Ours)	**89.4**	**82.8**	**86.0**	87.8	**81.6**	**84.7**	**89.6**	**82.4**	**85.9**	**88.4**	**84.6**	**86.5**

one more 1×1 convolution layer with BatchNorm and ReLU layers, whose total parameter volume is 264K, before the detection head module of baseline. Experimental results are shown in TABLE V and we can draw the following conclusion: implementing more parameters based on baseline deteriorates the detection performance of baseline in Total-Text [2], SCUT-CTW1500 [3] and ICDAR-2015 [36], which demonstrates that increasing parameters directly may not bring performance improvement. Therefore, above experiments verify that the improvement of proposed RCM and PFM comes from the superior module design.

E. Comparison with State-of-the-Arts

1) Comparison on Curved Text Detection Datasets: In this section, we evaluate the performance of our proposed model and compare with recent state-of-the-art methods on popular curved text detection benchmarks, Total-Text [2] and SCUT-CTW1500 [3].

As shown in TABLE VI, our RP-Text outperforms existing state-of-the-art models in F-measure on both Total-Text and SCUT-CTW1500 datasets. For Total-Text dataset, compared with our re-implemented PAN [5] baseline, our Region Context Module and Progressive Fusion Module bring 1.8% improvement, which is a significant progress in recent

TABLE VI
THE SINGLE-SCALE PERFORMANCES ON CURVED TEXT DATASETS. † DENOTES OUR RE-IMPLEMENTED PAN [5] MODEL. OUR RP-TEXT ACHIEVES STATE-OF-THE-ART PERFORMANCE ON BOTH TOTAL-TEXT AND SCUT-CTW1500 DATASETS.

Dataset		Total-Text [2]			SCUT-CTW1500 [3]		
Methodology	Venue	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
TextSnake [20]	ECCV-2018	82.7	74.5	78.4	67.9	**85.3**	75.6
SPCNet [9]	AAAI-2019	83.0	82.8	82.9	-	-	-
CSE [47]	CVPR-2019	81.4	79.7	80.2	78.7	76.1	77.4
PSENet [18]	CVPR-2019	84.0	78.0	80.9	84.8	79.7	82.2
LOMO [7]	CVPR-2019	88.6	75.7	81.6	**89.2**	69.6	78.4
CRAFT [17]	CVPR-2019	87.6	79.9	83.6	86.0	81.1	83.5
PAN-640 [5]	ICCV-2019	89.3	81.0	85.0	86.4	81.2	83.7
Boundary [48]	AAAI-2020	85.2	83.5	84.3	-	-	-
DB-R50-800 [16]	AAAI-2020	87.1	82.5	84.7	86.9	80.2	83.4
ContourNet [6]	CVPR-2020	86.9	83.9	85.4	83.7	84.1	83.9
DRRG [15]	CVPR-2020	86.5	**84.9**	85.7	85.9	83.0	84.5
PAN [5] †	ICCV-2019	88.4	80.3	84.2	88.0	80.1	83.9
RP-Text (Ours)	-	**89.4**	82.8	**86.0**	87.8	81.6	**84.7**

TABLE VII
THE SINGLE-SCALE PERFORMANCES ON THE QUADRILATERAL TEXT DATASETS. † DENOTES OUR RE-IMPLEMENTED PAN [5] MODEL. OUR RP-TEXT ACHIEVES STATE-OF-THE-ART PERFORMANCE ON MSRA-TD500 DATASET AND TOP PERFORMANCE ON ICDAR-2015.

Dataset		ICDAR-2015 [36]			MSRA-TD500 [4]		
Methodology	Venue	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
EAST [23]	CVPR-2017	83.6	73.5	78.2	87.3	67.4	76.1
SegLink [10]	CVPR-2017	73.1	76.8	75.0	86.0	70.0	77.0
RRPN [26]	TMM-2018	84.0	77.0	80.0	82.0	69.0	75.0
PixelLink [22]	AAAI-2018	85.5	82.0	83.7	83.0	73.2	77.8
Lyu et al. [21]	CVPR-2018	**94.1**	70.7	80.7	87.6	76.2	81.5
TextSnake [20]	ECCV-2018	84.9	80.4	82.6	83.2	73.9	78.3
SPCNet [9]	AAAI-2019	88.7	85.8	87.2	-	-	-
PSENet [18]	CVPR-2019	86.9	84.5	85.7	-	-	-
CRAFT [17]	CVPR-2019	89.8	84.3	86.9	88.2	78.2	82.9
PAN-736 [5]	ICCV-2019	86.6	79.7	83.0	85.7	83.4	84.5
Boundary [48]	AAAI-2020	88.1	82.2	85.0	-	-	-
DB-R50 [16]	AAAI-2020	91.8	83.2	**87.3**	**91.5**	79.2	84.9
ContourNet [6]	CVPR-2020	87.6	**86.1**	86.9	-	-	-
DRRG [15]	CVPR-2020	88.5	84.7	86.6	88.1	82.3	85.1
PAN [5] †	ICCV-2019	88.3	81.1	84.5	87.5	83.0	85.2
RP-Text (Ours)	-	89.6	82.4	85.9	88.4	**84.6**	**86.5**

scene text detection models. For SCUT-CTW1500 dataset, our RP-Text achieves new record 84.7% in F-measure with 0.8% improvement on baseline model. The superior performances on curved text detection benchmarks clearly demonstrate the effectiveness of our Region Context Module and Progressive Fusion Module to extract and fuse multi-scale text-related features for arbitrary-shaped text regions.

2) Comparison on Quadrilateral Text Detection Datasets: In this section, we further demonstrate the effectiveness of our RP-Text on quadrilateral text detection benchmarks, ICDAR-2015 [36] and MSRA-TD500 [4].

TABLE VIII
LEXICON-FREE END-TO-END TEXT SPOTTING COMPARISONS BETWEEN ABCNET BASELINE AND THAT WITH OUR RP-TEXT.

Dataset	Total-Text [2]		
Methodology	P (%)	R (%)	F (%)
Baseline	67.8	64.8	66.3
RP-Text (Ours)	**70.0**	**66.0**	**67.9**

Experimental results are shown in TABLE VII. For ICDAR-2015 dataset, input size of short side is the main difference between original PAN [5] (736) and our re-implement baseline model (1152). By enlarging the short side of input images from 736 to 1152, the performance improves 1.5% in F-measure (from 83.0 to 84.5). Moreover, due to RCM and PFM, our RP-Text has 1.4% improvement than baseline model and achieves top performance on ICDAR-2015. For MSRA-TD500 dataset, our RP-Text greatly outperforms current state-of-the-art text detection methodologies with a large margin (1.4% in F-measure).

TABLE IX
TEXT DETECTION COMPARISONS BETWEEN EAST BASELINE AND THAT WITH OUR RP-TEXT.

Dataset	ICDAR-2015 [36]		
Methodology	P (%)	R (%)	F (%)
Baseline	81.9	82.1	82.0
RP-Text (Ours)	**82.2**	**84.0**	**83.1**

TABLE X
SPEED ANALYSIS FOR OUR RP-TEXT. ALTHOUGH INCORPORATING RCM AND PFM, OUR RP-TEXT ONLY INCREASES LIMITED COMPUTATION COST AND KEEPS REAL-TIME INFERENCE FOR SCENE TEXT DETECTION.

Dataset	short size	PAN [5]†	RP-Text(Ours)	ms
		FPS		
Total-Text [2]	640	36.8	31.7	+4.37
SCUT-CTW1500 [3]	700	28.7	23.8	+7.17
ICDAR-2015 [36]	1152	14.9	13.7	+5.88
MSRA-TD500 [4]	736	36.9	27.3	+9.53

F. Plug-and-play Characteristic

To verify the plug-and-play characteristic of our model, we conduct the performance comparisons on ABCNet [28] and EAST [23] baselines. (1) ABCNet [28] provides experiments on curved text detection datasets such as Total-Text [2] in official implementation, therefore we implement this model as our first baseline and validate the end-to-end performance on Total-Text [2]. (2) Since EAST [23] only provides experiments on straight text detection datasets such as ICDAR-2015 [36], we utilize EAST text detector as our second baseline and validate the detection performance on ICDAR-2015 [36].

Performances of RCM and PFM with above baselines[2] additionally demonstrate the effectiveness of our proposed modules, which are shown in TABLE VIII and IX. Above experimental results verify that our plug-and-play RCM and PFM can boost the performances in widely-used baselines.

G. Visualization for Text Detection

We provide some visualization comparisons between PAN [5] and our proposed RP-Text in Fig. 9. According to first column, PAN erroneously recognizes key-like decorations and vertical lights as text regions. Our RP-Text, however, can better filter out mentioned background signals when detecting text instances from natural scenes. Besides, the second column indicates that our model can effectively extract text-related contextual information to separate text regions from complicated background and effectively recover the missing text region in natural scenes, especially for small and inconspicuous text regions. Moreover, The third and fourth columns of Fig. 9 show that comparing with PAN which detects incomplete text regions or missing some text regions,

[2]The baseline results from TABLE VIII and IX are our reimplemented results. The codebases we use are from:
https://github.com/aim-uofa/AdelaiDet/blob/master/configs/BAText
https://github.com/SakuraRiven/EAST.

our RP-Text is robust for locating text instances with large scale variance.

H. Speed Analysis

In this section, we study the inference time for our proposed method. We test our model and PAN [5] baseline on one single RTX-2080Ti GPU and Intel(R) Xeon(R) CPU E5-2650 v2 with 2.60GHz. Experimental results in TABLE X clarify that our method takes extra computation time from 4.37 ms to 9.53 ms per image and keep real-time inference for scene text detection.

V. Conclusion

In this paper, we analyse the weakness of current segmentation-based scene text detection methods and propose the Region Context Module and Progressive Fusion Module for accurately locating text instances. The Region Context Module (RCM) is developed to extract text-related scale-specific information based on semantic response in local regions while Progressive Fusion Module (PFM) utilizes multiple RCM branches to collect multi-scale contextual information. Our RCM and PFM are plug-and-play modules and we incorporate them into existing scene text detection platforms. Experimental results demonstrate the effectiveness of our proposed modules and we achieve state-of-the-art performances on Total-Text, SCUT-CTW1500 and MSRA-TD500 datasets.

References

[1] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[2] C. K. Ch'ng, C. S. Chan, and C. Liu, "Total-text: Towards orientation robustness in scene text detection," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 23, pp. 31–52, 2020.

[3] Y. Liu, L. Jin, S. Zhang, C. Luo, and S. Zhang, "Curved scene text detection via transverse and longitudinal sequence connection," *Pattern Recognition*, vol. 90, pp. 337–345, 2019.

[4] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 1083–1090.

[5] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu, G. Yu, and C. Shen, "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[6] Y. Wang, H. Xie, Z.-J. Zha, M. Xing, Z. Fu, and Y. Zhang, "Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[7] C. Zhang, B. Liang, Z. Huang, M. En, J. Han, E. Ding, and X. Ding, "Look more than once: An accurate detector for text of arbitrary shapes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[8] X. Wang, Y. Jiang, Z. Luo, C.-L. Liu, H. Choi, and S. Kim, "Arbitrary shape scene text detection with adaptive text region representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6449–6458.

[9] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao, and G. Li, "Scene text detection with supervised pyramid context network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9038–9045.

[10] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[11] 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.

[12] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," *arXiv preprint arXiv:1611.06779*, 2016.

[13] Y. Liu and L. Jin, "Deep matching prior network: Toward tighter multi-oriented text detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[14] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *European conference on computer vision*. Springer, 2016, pp. 56–72.

[15] S.-X. Zhang, X. Zhu, J.-B. Hou, C. Liu, C. Yang, H. Wang, and X.-C. Yin, "Deep relational reasoning graph network for arbitrary shape text detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[16] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time scene text detection with differentiable binarization." in *AAAI*, 2020, pp. 11 474–11 481.

[17] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[18] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, and S. Shao, "Shape robust text detection with progressive scale expansion network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[19] C. Xue, S. Lu, and W. Zhang, "Msr: Multi-scale shape regression for scene text detection," *arXiv preprint arXiv:1901.02596*, 2019.

[20] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "Textsnake: A flexible representation for detecting text of arbitrary shapes," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[21] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[22] D. Deng, H. Liu, X. Li, and D. Cai, "Pixellink: Detecting scene text via instance segmentation," *CoRR*, vol. abs/1801.01315, 2018. [Online]. Available: http://arxiv.org/abs/1801.01315

[23] 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 the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] 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*, vol. 20, no. 11, pp. 3111–3122, 2018.

[27] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[28] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, and L. Wang, "Abcnet: Real-time scene text spotting with adaptive bezier-curve network," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[29] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[30] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[31] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[32] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[33] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*, 2017, pp. 933–941.

[34] J. Fu, J. Liu, Y. Wang, Y. Li, Y. Bao, J. Tang, and H. Lu, "Adaptive context network for scene parsing," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 6748–6757.

[35] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[36] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny, "Icdar 2015 competition on robust reading," in *13th IAPR International Conference on Document Analysis and Recognition, ICDAR 2015 - Conference Proceedings*, ser. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. United States: IEEE Computer Society, Nov. 2015, pp. 1156–1160, 13th International Conference on Document Analysis and Recognition, ICDAR 2015 ; Conference date: 23-08-2015 Through 26-08-2015.

[37] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[38] C. Yao, X. Bai, and W. Liu, "A unified framework for multioriented text detection and recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4737–4749, 2014.

[39] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *CoRR*, vol. abs/1606.04797, 2016. [Online]. Available: http://arxiv.org/abs/1606.04797

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[41] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 761–769.

[42] C. Wolf and J.-M. Jolion, "Extraction and Recognition of Artificial Text in Multimedia Documents," *Pattern Analysis and Applications*, vol. 6, no. 4, pp. 309–326, 2003.

[43] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[44] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[45] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[46] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[47] Z. Liu, G. Lin, S. Yang, F. Liu, W. Lin, and W. L. Goh, "Towards robust curve text detection with conditional spatial expansion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

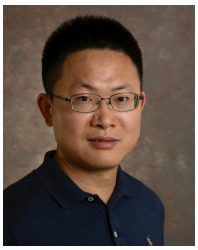
[48] H. Wang, P. Lu, H. Zhang, M. Yang, X. Bai, Y. Xu, M. He, Y. Wang, and W. Liu, "All you need is boundary: Toward arbitrary-shaped text spotting," *arXiv preprint arXiv:1911.09550*, 2019.

Qitong Wang received the B.S. in software engineering from Wuhan University of Technology, China in 2018, and M.S. in computer science from Boston University, USA in 2020. Since 2021, he has been with the Department of Computer Information Sciences, University of Delaware, USA, as a Ph.D. student. His research interests include computer vision and machine learning.

Bin Fu is currently an Assistant Research Fellow with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He received the Ph.D. degree from the University of Hong Kong, in 2018 and B.E. degree from Lanzhou University, in 2014. His research interests include semantic segmentation and scene text recognition.

Ming Li received the B.S. degree from Xi'an Jiaotong University, China, in 2020 and served as a Research Assistant in ShenZhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, from 2019 to 2021. He is currently a Master student at Texas A&M university, US. His research interests include scene text detection, scene text recognition and Natural Language Processing.

Junjun He is currently a Research Assistant with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences and Shanghai AI Laboratory. His research interests include computer vision and medical image computing, especially on dense prediction, multi-view learning, multi-modal learning and efficient model design. He has published more than 10 papers in international journals and conferences, including T-PAMI, TMI, CVPR, ICCV, ECCV, MICCAI, ISBI etc.



Xi Peng is an assistant professor of the Department of Computer and Information Sciences and a resident faculty of the Data Science Institute, both at the University of Delaware, Newark, DE, USA, where he is directing the Deep-REAL research group. His current research interest is developing flexible, reliable, and explainable machine learning methods, upon which computer vision, biomechanics, and geoinformatics can advance synergistically. He received the B.E. degree from Beihang University, Beijing, China, in 2008, the M.E. degrees from Chinese Academy of Sciences, Beijing, China, in 2011, and the Ph.D. degree from Rutgers University, NJ, USA, in 2018.

Yu Qiao is a professor with the Shenzhen Institutes of Advanced Technology (SIAT), the Chinese Academy of Science and Shanghai AI Laboratory. His research interests include computer vision, deep learning, and bioinformation. He has published more than 240 papers in international journals and conferences, including T-PAMI, IJCV, T-IP, T-SP, CVPR, ICCV etc. His H-index is 67, with 28,000 citations in Google scholar. He is a recipient of the distinguished paper award in AAAI 2021. His group achieved the first runner-up at the ImageNet Large Scale Visual Recognition Challenge 2015 in scene recognition, and the winner at the ActivityNet Large Scale Activity Recognition Challenge 2016 in video classification. He served as the program chair of IEEE ICIST 2014.