R-Net: A Relationship Network for Efficient and Accurate Scene Text Detection

Yuxin Wang, Hongtao Xie, Zhengjun Zha, Youliang Tian, Zilong Fu, and Yongdong Zhang, *Senior Member, IEEE*

***Abstract*—This paper introduces a novel bi-directional convolutional framework to cope with the large-variance scale problem in scene text detection. Due to the lack of scale normalization in recent CNN-based methods, text instances with large-variance scale are activated inconsistently in feature maps, which makes it hard for CNN-based methods to accurately locate multi-size text instances. Thus, we propose the relationship network (R-Net) that maps multi-scale convolutional features to a scale-invariant space to obtain consistent activation of multi-size text instances. Firstly, we implement an FPN-like backbone with a Spatial Relationship Module (SPM) to extract multi-scale features with powerful spatial semantics. Then, a Scale Relationship Module (SRM) constructed on feature pyramid propagates contextual scale information in sequential features through a bi-directional convolutional operation. SRM supplements the multi-scale information in different feature maps to obtain consistent activation of multi-size text instances. Compared with previous approaches, R-Net effectively handles the large-variance scale problem without complicated post processing and complex hand-crafted hyperparameter setting. Extensive experiments conducted on several benchmarks verify that our R-Net obtains state-of-the-art performance on both accuracy and efficiency. More specifically, R-Net achieves an F-measure of 85.6% at 21.4 frames/s and an F-measure of 81.7% at 11.8 frames/s for ICDAR 2015 and MSRA-TD500 datasets respectively, which is the latest SOTA. The code is available on *https://github.com/wangyuxin87/R-Net*.**

***Index Terms*—Scene text detection, large-variance scale, convolutional neural network (CNN).**

I. Introduction

EXTRACTING and recognizing textual information in the wild have become increasingly important and popular in recent years due to its significant value in practical applications [1]–[7]. Scene text detection, playing a critical role in end-to-end text reading task, is one of the main bottlenecks in recognition quality.

(a) Cropped input (b) EAST (c) R-Net

Fig. 1. We visualize the feature maps from parallel method EAST [18] and the proposed method. Both models use VGG16 as their basic network. (a) Identical cropped image is used as input for both models. (b) The convolutional features of the last decoding layer in EAST. Multi-size texts can not be activated consistently and obviously. (c) Convolutional features from our R-Net. Our method obtains features with consistent activation of multi-size texts.

Benefiting from the development of general object detection algorithms [8], [9], scene text detection has achieved great improvement in recent years [10]–[13]. However, most of them suffer from the large-variance scale problem, which makes it hard for recent CNN-based methods to accurately locate multi-size text instances. As feature pyramid facilitates the accurate detection of multi-size objects [3], [14], [15], some recent approaches combine predictions from pyramid features to handle this problem. Lyu *et al.* [16] predict corner points of text instances from multi-scale features, then Non-maximum Suppression (NMS) is used for the reduction of redundant boxes. Liao *et al.* [17] detect multi-size texts by combining predictions from rotation-sensitive and rotation-invariant branches in feature pyramid. Though these methods are able to handle texts with large-variance scales, the *multi-forward* predictions obtained by combining predictions from feature pyramid will cause amounts of redundant bounding boxes, which results in large time consumption and is difficult for practical application.

Different from previous methods, we regard the large-variance scale problem as the issue of inconsistent activation of multi-size texts in feature maps. As shown in Fig. 1(b), the multi-size texts can not be activated simultaneously and observably in the same feature map. This inconsistence is caused by two conditions: 1) pooling operations lose the accurate geometry information of small texts. 2) The large texts can not receive enough representation in shallow stages. Since it is difficult to balance

Manuscript received October 15, 2019; revised February 26, 2020 and April 10, 2020; accepted May 5, 2020. Date of publication May 19, 2020; date of current version April 23, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFC0820600, in part by the National Nature Science Foundation of China under Grants 61525206 and 61771468, and in part by the Youth Innovation Promotion Association Chinese Academy of Sciences under Grant 2017209. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Lamberto Ballan. *(Corresponding authors: Hongtao Xie; Yongdong Zhang.)*

Yuxin Wang, Hongtao Xie, Zhengjun Zha, Zilong Fu, and Yongdong Zhang are with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: wangyx58@mail.ustc.edu.cn; htxie@ustc.edu.cn; zhazj@ustc.edu.cn; jeromef@mail.ustc.edu.cn; zhyd73@ustc.edu.cn).

Youliang Tian is with the Guizhou University, Guiyang 550025, China (e-mail: yltian@gzu.edu.cn).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2020.2995290

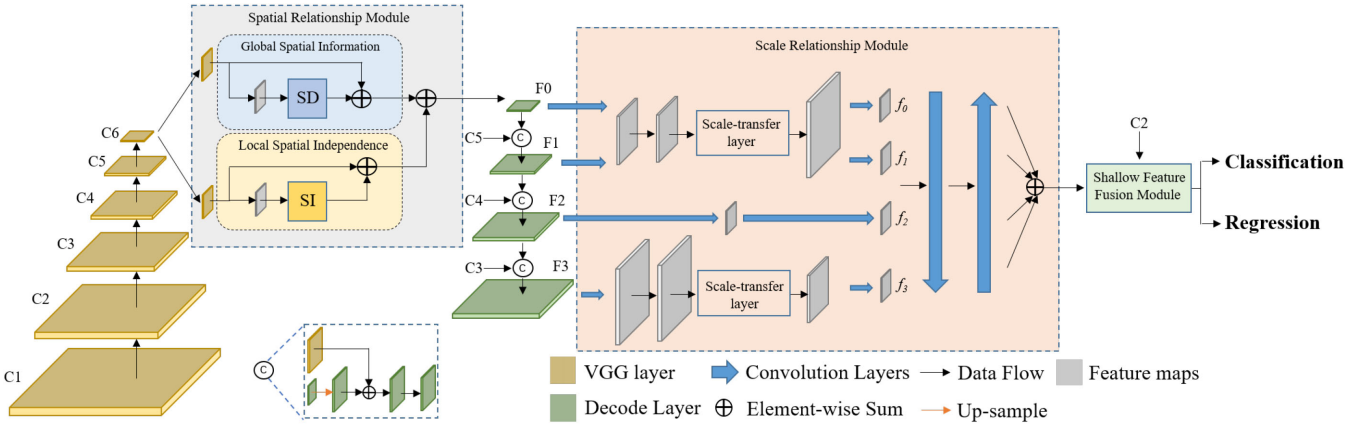


Fig. 2. The architecture of R-Net. The network contains four parts: backbone, Spatial Relationship Module (SPM), Scale Relationship Module (SRM) and Shallow Feature Fusion Module (SFFM). The backbone is adapted from VGG16. SRM is built on feature pyramid to pass contextual scale information. Output features from SFFM are used for bounding box prediction (classification and regression).

the semantics between texts with different scales in feature maps, accurately detecting scene texts with large-variance scales is still challenging. Directly fusing pyramid feature maps is the most straightforward approach to generate features with rich scale information. However, it has two shortcomings (2-shorts): 1) multi-scale feature maps have different resolutions, which is difficult for direct addition. 2) It would limit the performance of the network to a certain degree due to the heterogeneity of features.

In this paper, we propose a novelly efficient and accurate bi-directional convolutional framework called R-Net to effectively handle the large-variance scale problem, which aims to map multi-scale convolutional features to a scale-invariant space and achieves accurate detection results from **Only One** convolutional layer (*single-forward* predictions). To effectively handle the 2-shorts, we propose a novel Scale Relationship Module (SRM) constructed on pyramid features for joint resolution normalization and multi-scale feature aggregation. As shown in Fig. 2, SRM contains two parts: scale-transfer layer and bi-directional convolutional operation. The scale-transfer layer efficiently normalizes multi-scale features to a unified resolution. Then bi-directional convolutional operation propagates contextual scale information in sequential normalized feature maps to supplement multi-scale information in different convolutional layers and generates features with consistent activation of multi-size texts. To further enhance the representation of scene texts in complex background, we propose a Spatial Relationship Module (SPM) in the first decoding layer to enhance the spatial semantics by jointly considering the long-range dependencies and local independencies. Compared with previous methods, the proposed R-Net effectively addresses the large-variance scale problem without complex hand-crafted hyperparameter setting and complicated post processing.

Different from DSRN [19] which uses a similar approach for handling the large-variance scale problem, the proposed R-Net benefits in following two aspects: 1) we study the relationship between multi-scale features in a smaller resolution (1/8 of input image), which can further improve the efficiency of our method. 2) A new scale transfer layer is proposed to map the large-scale feature maps to a smaller one, which is proved to effectively maintain the geometrical characteristics of small-size texts and improve the detection performance (see in Table VII). Benefiting from these two aspects, our R-Net can obtain accurate detection results in large-variance scales with high efficiency.

The contributions of this paper are following:

- We propose a Scale Relationship Module to essentially handle the large-variance scale problem by mapping multi-scale features to a scale invariant space to obtain features with consistent activation of multi-scale texts.
- A Spatial Relationship Module is designed to enhance spatial semantics by jointly considering the long-range dependencies and local independencies.
- The proposed SPM and SRM can be easily embedded into existing methods, boosting the performance without obvious speed sacrifice.
- Our R-Net achieves state-of-the-art results in both accuracy and efficiency. Specifically, R-Net achieves an F-measure of **85.6%** at **21.4 frames/s** and an F-measure of **81.7%** at **11.8 frames/s** on ICDAR 2015 dataset and MSRA-TD500 dataset respectively.

The remainder of this paper is organized as follows. Section II introduces the related works. In Section III, we discuss the proposed R-Net in detail. In Section IV, we demonstrate the quantitative studies on four datasets. Finally, we conclude our work in Section V.

II. Related Work

As deep learning becomes the most promising machine learning tool [20], [21], scene text detection has achieved remarkable improvement in recent years with many solutions proposed [16], [18], [22]. In this paper, we summarize scene text detection methods into two categories: proposal-based methods and proposal-free methods. We will introduce them respectively and elaborate the difference between these methods and ours.

A. Proposal-Based Methods

Benefiting from the development in recent object detection techniques [8], [9], [23], many methods adopt similar approaches to achieve accurate text localization. Based

on SSD [15], Liao *et al.* [17] adopt ARF [24] to generate rotation-sensitive features to cope with the multi-oriented text instances. To enhance the representation ability of multi-size texts, SSTD [11] develops a hierarchical inception module to aggregate multi-scale information from pyramid features. TextBoxes++ [12] represents arbitrary-oriented texts with quadrilaterals. Wang *et al.* [25] use a recurrent neural network (RNN) to iteratively predict pairs of boundary points for the representation of text regions. Due to the experiential setting in label generation (point pair will be removed when the ratio of modified area to original area is less than 0.93) and iterative operation in RNN, [25] is sensitive to the hyper-parameter settings and results in large time consumption. FOTS [26] uses a RoIRotate algorithm to transform oriented feature regions to axis-aligned feature maps and decodes text bounding boxes from generated RoIs. Liu *et al.* [27] consider scene text detection as a region expansion problem and propose a conditional prediction process to retrieve the instance level text region by expanding outwards from the seed.

Although proposal-based methods achieve promising results in many situations, they also have three shortcomings. Firstly, many proposal-based methods rely on one-step regression on various layers (one-stage approaches [11], [17]), which is inaccurate and inefficient in some challenging scenarios. Besides, the detection performance is sensitive to the anchor settings. Secondly, the threshold setting is highly adversarial. Though higher threshold lets positive anchors contain less background information, it reduces the number of positive anchors and may cause imbalance problem during training stage. In contrast, the lower values can generate more diverse positive proposals and provide a little incentive to reduce false positives. Thirdly, the large amount of anchors would increase the computation and degrade the efficiency of matching between ground truths and proposals. Compared with these proposal-based approaches, the proposed R-Net effectively handles the large-variance scale problem without complex hand-crafted hyperparameter setting (e.g. scale of anchors) and complicated post processing, which is a more robust approach for scene text detection and can obtain a better trade-off between the accuracy and efficiency.

B. Proposal-Free Methods

Proposal-free methods can be further divided into segmentation based methods and direct regression based methods.

1) Segmentation Based Methods: Segmentation based methods have been popular in recent years, these methods represent complex text instances with pixel-level prediction. Inspired by popular segmentation methods [28], [29], some algorithms rebuild texts from segmentation maps [30]–[32]. As the common geometric feature of text resembles a snake, TextSnake [30] intuitively uses several disks to cover every text instance. PixelLink [31] predicts the links among every pixel and their neighbors, and the links are valid when both of the linked pixels belong to the same text instance. To split the close text instances, Wang *et al.* [32] firstly predict multiple segmentation results at certain scales, then a progressive scale expansion is used to cope with the easily confused area by gradually expanding contextual kernels with different scales. CRAFT [33] generates bounding boxes by combining character-wise predictions with learned affinity between adjacent character regions. However, the limitation of requirement of additional character-level annotations and the large time consumption of bounding box generation make it difficult for practical application. Liu *et al.* [34] propose a bottom-up approach to treat scene text detection as a graph clustering problem by performing Markov Clustering on the predicted node graph. Due to the missed geometrical information in the node map, [34] obtains unsatisfactory results in small-size text detection. Tian *et al.* [22] regard scene text localization as a instance segmentation task, they predict the embedding map and leverage a two-step clustering to split adjacent instances. However, due to the fairly complicated post processing, the segmentation based methods usually result in large time consumption.

2) Direct Regression Based Methods: Different from previous methods, DDR [35] and EAST [18] propose a new method called direct regression based method (DRBM), which directly predicts offsets from bounding box boundaries or vertexes to pixels. In addition, due to the abandonment of complex anchor setting and complicated post processing, DRBM is a more robust approach for scene text detection and can achieve a better trade-off between accuracy and efficiency.

However, simply concatenating features with multi-scale information limits the representing ability of large-variance scale texts due to the heterogeneity of features. It is difficult for DRBMs to balance semantics between texts with different scales in feature maps, which results in inconsistent activation of multi-size texts (Fig. 1(b)). Thus, we propose a novel approach to handle the large-variance scale problem by constructing a bi-directional convolutional operation on feature pyramid to map multi-scale features to a scale-invariant space. As shown in Fig. 1(c), our method obtains more consistent activation of multi-size texts and achieves a better balance of semantics between texts with different scales. The details of our method will be introduced in Section III.

C. Bi-Directional Operations

Bi-directional operations have been widely used in recent CNN-base methods for both detection and recognition tasks. Zeng *et al.* [36] integrate contextual visual cues of RoIs in two directions to learn their nonlinear relationships. He *et al.* [37] use a bi-directional approach to encode sequential context features. Different from these methods, our proposed method aims to map multi-scale convolutional features to a scale-invariant space and obtain consistent activation of multi-size texts. To be specific, the proposed method models the relationship between different features by passing the contextual scale information in sequential feature maps to supplement multi-scale information, which is a novel approach to address the large-variance scale problem compared with previous methods.

III. Proposed Method

In this section, we will introduce our R-Net in detail. The total architecture is illustrated in Fig. 2. R-Net contains four

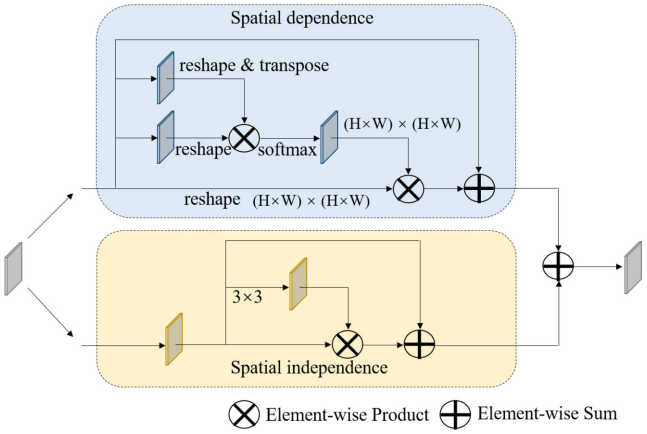


Fig. 3. The architecture of Spatial Relationship Module. Two branches in different color represents the pipeline of spatial dependency and spatial independency respectively.

parts: backbone, Spatial Relationship Module, Scale Relationship Module and Shallow Feature Fusion Module. The output features from Shallow Feature Fusion Module are used for text bounding box prediction.

A. Feature Extraction

The backbone of R-Net is adopted from a pre-trained VGG16 [38] and designed with following considerations: 1) the backbone must have enough capacity to represent the large variance of text scales. 2) Features should contain more contextual information for the robust representation. Inspired by FPN [14] which achieves the good performance on mentioned problems, we adopt a FPN-like architecture to extract features.

Particularly, we convert the fully connected layers in VGG16 [38] to Spatial Relationship Module (in Section 3.2). Then several extra convolutional layers (*F0*, *F1*, *F2* and *F3*) are stacked above SPM in a top-down pathway, and lateral connections [14] is implemented for generating high-level semantic features at all scales. Note that the extracted features from backbone are shared in our design, thus the scheme is parameter-efficient.

B. Spatial Relationship Module

The attention mechanism has been adopted in many recent works to select a focused location and boost the discriminative features. Long-range dependencies are certified to be necessary for object detection task [39]. Despite convolutional operation can capture long-range dependencies by stacking convolution layers, it suffers from inefficient computational ability and optimization difficulties. To solve this problem, Wang *et al.* [40] propose a non-local operation to efficiently capture long-range dependencies. Inspired by [40], we firstly implement SD branch to capture the long-range dependencies in feature maps (top branch in Fig. 3).

SD branch sums up the weighted features in a long-range approach and calculates the spatial dependencies by a dilated residual network. Firstly, we generate a spatial attention matrix to model the long-distance spatial relationship between any two pixels. Then the enhanced features are obtained through a matrix multiplication between original features and spatial attention matrix. Finally, the final representations reflecting long-range contexts are generated through an element-wise sum operation between the original features and enhanced features. For the practical implementation of SD branch, we refer to some implementation details from PAM in DANet [29].

$$Position_i = \frac{1}{N(x)} \sum_j f(x_i, x_j) g(x_j) \tag{1}$$

The outputs of SD branch are formulated by Equation (1), x is the input feature, $Position_i$ is the output feature in position i, f computes the dependencies between x_i and x_j, N computes the normalization factor. For each position i, Equation (1) calculates the dependencies between i and other positions, and the response to local correlation will be weaken when the resolution of feature map is large (16×16 in our method), which is sub-optimal for spatial semantics enhancement. Thus we propose a parallel SI branch to only enhance the local spatial independencies for learning local correlation which is proved also necessary for scene text detection [34].

As shown in the bottom of Fig. 3, SI branch filters the inputs with attention maps which are rich in locally discriminating information. The input feature maps are firstly processed by two 3×3 convolution layers, then element-wise product is implemented to filter inputs and highlight the local visual cues. Finally, we perform an element-wise sum operation to fuse spatial dependence and independence information from two branches.

The proposed SPM models the spatial dependencies and independencies simultaneously. The long-range dependencies can model the long-distance relationship between pixels, and the independence information can highlight the local visual cues for enhancement of local correlation. Considering the following two conditions: 1) representing texts with large-variance scale relies heavily on the deep features with strong semantic information; 2) efficiency of our method. We only construct SPM in the last encoding layer.

C. Scale Relationship Module

The accurate geometry prediction of small texts requires low-level information from early stage of convolutional layers, while the localization of large texts requires features from late stage with relatively low resolution [18]. However, due to the heterogeneity of features, simply concatenating features from different stages carrying multi-scale information would limit the performance of the network. Thus, different from previous methods, we design a Scale Relationship Module (SRM) to aggregate multi-scale information based on bi-directional convolutional operations. SRM contains two part: scale-transfer layer and bi-directional convolutional layer. These two parts will be elaborated sequentially.

Since it is difficult to aggregate features from different stages with various resolutions, inspired by [41], we firstly propose a scale-transfer layer to map multi-scale features to a unified resolution (1/8 of input image in this paper). The architecture of scale-transfer layer is shown in Fig. 4. We firstly use a 1×1 convolutional operation in channel matching layers to normalize output channels. This operation can also keep smoothness for features and enhance learning ability. Then the following

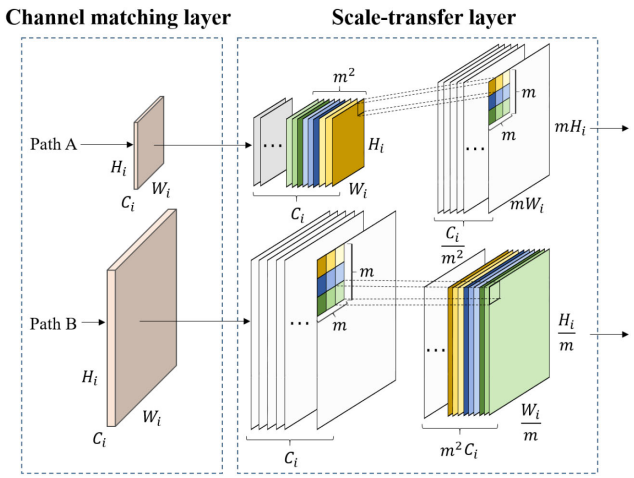


Fig. 4. The architecture of scale-transfer layer.

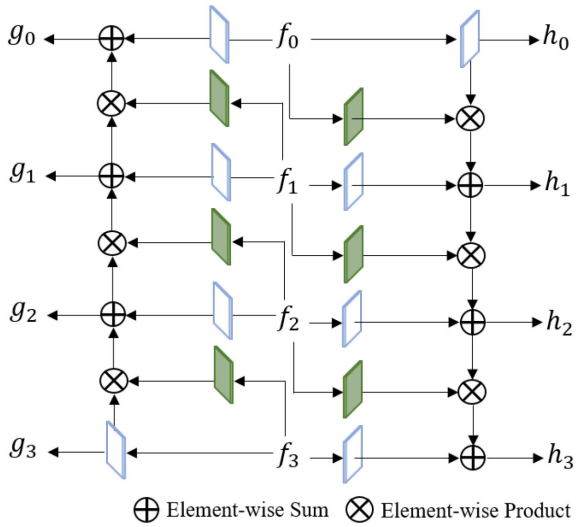


Fig. 5. The architecture of bi-directional convolutional layer.

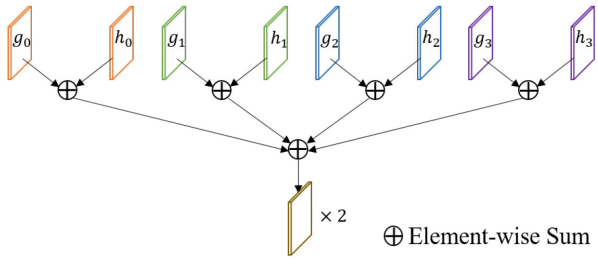


Fig. 6. The processing of feature fusion after bi-directional convolutional layer.

two-path scale-transfer operation is proposed to normalize the resolution of feature maps. As illustrated in Fig. 4, path A expands hight and width of feature maps simultaneously by compressing the channels. In contrast, path B leverages an inverse approach compared to path A which compresses the height and width of features to expand the channels. The dimensions of input tensor are $C_i \times H_i \times W_i$ $(i = 1, 2, 3)$ and those of outputs are $\frac{C_i}{m^2} \times m \times H_i \times m \times W_i$. We set $m = 4, 2, 1, 1/2$ for $F0, F1, F2, F3$ respectively. Note that the proposed scale-transfer operations contain no extra parameters except channel matching layers. Due to the fact that channels of features are usually related to stages, channel normalization should also be implemented following other resizing methods. Thus the proposed scale-transfer layer normalizes the resolution of feature maps with less additional calculating consumption compared with existing approaches, which keeps the efficiency of our network. In the end, the output features of the scale-transfer layer have the same dimension to be compatible.

After normalization, we develop a bi-directional convolutional operation to pass the contextual scale information in sequential feature maps and aggregate features from each step to generate features with rich scale information. The architecture of bi-directional convolutional layer is illustrated in Fig. 5.

The bi-directional convolutional layer is conducted on multi-stage layers to sequentially convolute the normalized features from scale-transfer layer. f_0, f_1, f_2 and f_3 are corresponding normalized features from F_0, F_1, F_2 and F_3 in Fig. 2 respectively. Sequential convolutional operation in the first direction starts from the first decoding layer (f_0) and ends at the last decoding layer (f_3), thus the sequential feature maps can receive large-scale information from small-size features, which is better for representing large-size texts. In contrast, sequential feature maps can also receive contextual small-scale information in the second direction, which helps locate small-size texts. The first output features are processed by a 3×3 convolutional layer in Equation (2). As illustrated in Equation (3), a single-channel feature map is normalized by sigmoid activation function. Equation (4) has the similar spirit to residual-connection [42], which is proposed to prevent the filtered feature from degradation. We integrate the attention map into previous feature map h_{i-1} via element-wise product, aiming to control the contextual scale information passing to the next stage. The midline features (h'_i) are fused with convoluted features from f_i as the final outputs (h_i).

$$h_0 = \sigma(w_0 \times f_0 + b_0) \tag{2}$$

$$att_i = sigmoid(w'_i \times f_{i-1} + b'_i) \tag{3}$$

$$h'_i = \sigma(w_{1i} \times h_{i-1} + b_{1i}) \odot att_i + \sigma(w_{2i} \times h_{i-1} + b_{2i}) \tag{4}$$

$$h_i = cat(h'_i\ ,\ \sigma(w^*{}_i \times f_i + b^*{}_i)) \tag{5}$$

Where σ denotes the ReLU operation, $\times$ denotes the convolution operation, cat is the concatenation operation and $\odot$ denotes element-wise product. In contrast, the operations in the second direction have the similar forms to the operations in the first direction. After the bi-directional convolutional operations, the output feature maps (g_i and h_i, $i = 0, 1, 2, 3$) have richer multi-scale and stronger semantic information than previous feature maps ($f_i, i = 0, 1, 2, 3$). In the end, we concatenate the output features and implement two sequential 3×3 convolutional layers to obtain better representation (Fig. 6).

D. Shallow Feature Fusion Module

Due to the fact that tiny and dense detections are existing with high probability in scene text detection task, following the spirit

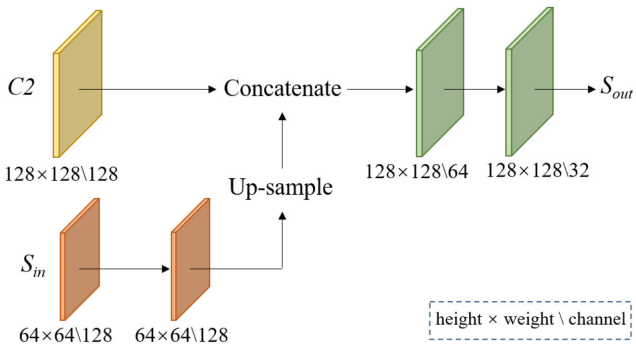


Fig. 7. The architecture of Shallow Feature Fusion Module.

of FPN [14], we introduce a Shallow Feature Fusion Module (SFFM) to expand resolution of feature maps and enhance the ability of tiny and dense detections (Fig. 7). Within SFFM, input features are firstly up-sampled to the same resolution as previous shallow feature maps ($C2$). Afterwards, shallow and up-scaled features are concatenated and followed by two 3×3 convolution layers. The proposed SFFM produces the final prediction S_{out} as:

$$S' = \sigma(w_{o1} \times (cat(up(S_{in}), C2)) + b_{o1}) \tag{6}$$

$$S_{out} = \sigma(w_{o2} \times S' + b_{o2}) \tag{7}$$

Where σ denotes the ReLU operation, $\times$ denotes the convolution operation, cat is the concatenation operation and up is the up-sample operation (bilinear interpolation is used in this paper).

E. Loss Function

The multi-task loss is formulated as:

$$L = L_{cls} + \lambda_{reg} L_{reg} \tag{8}$$

Where L_{cls} and L_{reg} denote classification loss and regression loss respectively. λ_{reg} is a hyperparameter to balance the two loss values and we set λ_{reg} to 1 in our experiment.

As for classification loss, we use Dice Loss to optimize L_{cls}. Dice Loss is constructed in (9). g_i and p_i denote pixels in ground truth and prediction respectively.

As for regression task, we choose the similar direct regression approach as RBOX regression in [18], which is invariant against scales of objects (Equation (10) (11)). Where P denotes the predicted box and G is the corresponding ground truth. θ' denotes the predicted angle and θ^* is the ground truth.

$$L_{cls} = 1 - \frac{2\sum_i g_i p_i}{\sum_i g_i + \sum_i p_i} \tag{9}$$

$$L_{loc} = -\log IoU(P, G) = -\log \frac{|P \cap G|}{|P \cup G|} \tag{10}$$

$$L_\theta = 1 - \cos(\theta' - \theta^*) \tag{11}$$

$$L_{reg} = L_{loc} + L_\theta \tag{12}$$

We combine L_{loc} with L_θ as the final regression loss in Equation (12). Distance from each pixel to 4 boundaries, value of angle and probability map are outputs for both training and inference.

IV. Experiment

In this Section, we evaluate the proposed method on four popular text detection benchmarks: ICADAR2015, MSRA-TD500, ICDAR2013 and MLT. Firstly, to demonstrate the effectiveness of SPM and SRM, we conduct several ablation studies. Then we investigate the influence of hyperparameters in these two proposed modules. In the end, we compare our R-Net with recent state-of-the-art methods on benchmark datasets to demonstrate our accuracy and efficiency.

A. Benchmark Datasets

1) SynthText: The SynthText [43] is a synthetically generated dataset created via blending rendered words with natural images which contains 800 k images. Following by recent methods [16], [22], we use this dataset to pre-train our model.

2) ICDAR 2013: The ICDAR 2013 dataset [44] contains 229 training images and 233 images for testing in different resolution. Particularly, this dataset only contains horizontal texts. In test stage, we set the θ to 0 to generate horizontal predictions.

3) ICDAR 2015: The ICDAR2015 is a dataset for incidental scene text detection proposed in the Challenge 4 of ICDAR 2015 Robust Reading Competition [45]. It includes 1500 images (1000 images for training and 500 images for testing) with annotations labeled as 4 vertices. In training stage, we fit a rotated rectangle with the minimum area. Different from previous dataset for only horizontal texts, this benchmark is proposed for evaluating texts of different scales, ambiguities, resolutions, perspectives, and arbitrary orientation.

4) MLT: The MLT [46] is a dataset proposed on ICDAR2017 Competition for focusing on multi-script, multi-oriented and multi-lingual aspects of texts. It consists of 9000 for training (7200 training images and 1800 validation images) and 9000 for testing.

5) MSRA-TD500: The MSRA-TD500 [47] is a dataset for detecting multi-lingual and arbitrary-oriented long text lines. It contains 500 images (300 images for training and 200 images for testing) which consists of Arabic numbers and English letters of different fonts, and the labels are all in line-level. Since the size of training images is too small to learn a deep network, 400 images from HUST are also used in training stage.

6) HUST: The HUST [47] is a dataset containing 400 images, which includes and Arabic numbers of different fonts with text line level labels. We combine this dataset with MSRA-TD500 for training.

B. Implementation Details

The backbone of our network is VGG16 [38]. The channels during up-sample pipeline are set to 512, 128, 64, 32 respectively. For data augmentation, we first randomly rescale the height of input images with ratio from 0.8 to 1.2, then rotate it with angle from -10 to 10. Finally, the training images are generated by randomly cropping 512×512 patches from the

TABLE I
THE PERFORMANCE GAIN OF PROPOSED SPM AND SRM

SPM	SRM	R	P	F
×	×	80.2	86.2	83.1
✓	×	80.4	87.7	83.9
×	✓	82.0	86.5	84.2
✓	✓	**82.5**	**87.9**	**85.1**

TABLE II
PERFORMANCE GAIN OF SPATIAL RELATIONSHIP MODULE. SI AND SD ARE SHORT FOR SPATIAL INDEPENDENCY, AND SPATIAL DEPENDENCY BRANCH

Algorithm	SI	SD	R	P	F
Baseline	-	-	80.9	86.6	83.7
SRM + SPM	-	512	82.0	86.8	84.4
SRM + SPM	128	128	81.4	86.7	84.0
SRM + SPM	128	512	**83.1**	86.3	84.7
SRM + SPM	512	128	82.7	87.0	84.8
SRM + SPM	512	512	81.8	**88.8**	**85.2**

rotated images. Our proposed network is trained end-to-end on 2 NVIDIA TITANX Pascal GPUs using Adam optimizer with batch size 16. The batch normalization is also used in our implementation. We use initial learning rate 1e-4 to pre-train all models on SynthText, and then finetune them on corresponding tasks. Specifically, because of the limited amount of Chinese samples in MSRA-TD500, we firstly pretrain the model on 4 k samples from [48], then finetune it on corresponding dataset.

During inference, we test our method on 1 NVIDIA TITANX Pascal GPU. There are six outputs, e.g. four distances to corresponding sides, one angle map and one probability map. We reestablish boxes of texts with these six maps and use NMS to reduce the redundant boxes. Particularly, we only evaluate our model at single scale for all datasets, due to that the settings of multi-scale testing in public methods are quite different from each other, which is hard for fair comparison.

C. Ablation Studies

Several ablation experiments are conducted on ICDAR2015 [45], ICDAR2013 [44] and MSRA-TD500 [47] to analyze R-Net. We first conduct the ablation study one by one to show the effectiveness of each module. Then we study the hyper-parameters in SPM and SRM respectively.

1) Effectiveness of SPM and SRM: To show the effectiveness of each proposed module, we implement SRM and SPM one by one in the baseline network. We fix the channel in SRM to 256 and set the channel in both SI and SD branches to 512. We firstly pre-train all the model on a part of SynthText [43] (about 200 k images), and then finetune them on ICDAR2015 dataset. As shown in Table I, the proposed modules improve the performance of baseline model step by step. It is worth mentioning that, implemented with both SPM and SRM, the improvement significantly increases to 2.3%, 1.7% and 2% in precision, recall and F-measure respectively.

2) Spatial Relationship Module: **The ablation study about SI and SD branches.** An evaluation of SD and SI branch in SPM on ICDAR2015 is shown in Table II. We firstly pre-train all the model on the whole SynthText [43] for 1 epoch, and then finetune them on ICDAR2015. We use VGG16 [38] as the backbone, and fix the channel in SRM to 128. The baseline model is constructed with only SRM. As can be seen in Table II, the improvement obtained by SPM is incremental. Compared with baseline model, SPM achieves 85.2% in F-measure surpassing baseline model by 1.5 %. Further analysing of results in Table II give us the following insights: 1) the number of channel varies the capacity of SRM, which can help network capture powerful spatial dependencies and independencies. 2) SI branch obtains 0.3% and 0.8% improvement in F-measure when we set the number of channel in SI branch to 128 and 512 respectively, which proves our standpoint and demonstrates the local correlation is also important for accurate detection. 3) Spatial dependencies and independencies are interdependent for performance gain. We set SI and SD to 512 simultaneously in our final experiments.

TABLE III
THE COMPARISONS BETWEEN SPM AND RECENT ATTENTION MODULES ON ICDAR2015

Methods	R	P	F
SelectionGAN[49]	82.4	87.6	84.9
DANet [29]	82.0	87.2	84.5
SPM	**82.8**	**88.7**	**85.6**

Comparisons with other attention modules. To further evaluate the effectiveness of the proposed SPM in our method, we conduct several ablation studies to compare SPM with other attention modules proposed in [29], [49]. We refer to the implementation details in their released code. We firstly pre-train all the models on the whole SynthText [43] for 1 epoch, and then finetune them on ICDAR2015 dataset. All the models are implemented with SRM (256 channels). The results are shown in Table III. SelectionGAN [49] firstly uses a Multi-scale Spatial Pooling operation to calculate the attention maps in different sizes and then concatenates the filtered features in different branches. Finally, Multi-channel Attention Selection is used to refine the features. Due to the limitation of receptive field of convolutional kernels in [49], our method takes advantage in capturing the long-range dependencies by SPM and obtains 0.4%, 1.1% and 0.7% improvement in recall, precision and F-measure respectively compared with [46]. DANet [29] uses two parallel branches to calculate the long-range dependencies and channel-wise distinctive features respectively. The channel-wise attention operation is proposed to model channel interdependencies. Different from DANet, our SPM is proposed to only focus on spatial information (*e.g.* long-range dependencies and local independencies), which is designed to be more consistent. As discussed in Section. III B, the proposed SI branch supplements the ignored local correlation and helps SPM capture richer spatial information. Compared with DANet, our SPM obtains 0.8%, 1.5% and 1.1% improvement in recall, precision and F-measure respectively.

3) Scale Relationship Module: **The evaluation of bi-directional operation.** To demonstrate the effectiveness of SRM, we train two baseline models shown in Table IV. One baseline model (-) is trained without Scale Relationship

TABLE IV
PERFORMANCE GAIN OF SCALE RELATIONSHIP MODULE. † AND * MEANS EXPERIMENTS ON MSRA-TD500 AND ICDAR2015 RESPECTIVELY

Algorithm	R	P	F	FPS
Baseline(-)†	54.1	75.1	62.0	**16.1**
Baseline(UCO)†	67.0	84.7	74.8	15.4
SRM†	**71.2**	**87.6**	**78.5**	13.3
Baseline(-)*	61.2	70.0	65.7	**12.1**
Baseline(UCO)*	77.7	82.6	80.0	10.8
SRM*	**79.6**	**83.2**	**81.4**	8.8

TABLE V
THE COMPARISONS BETWEEN FPN AND SRM

Methods	R	P	F
Baseline	80.2	86.2	83.1
Baseline+ FPN	81.1	86.2	83.6
Baseline+ SRM	**82.0**	**86.5**	**84.2**

Module, and another baseline model (UCO) only uses unidirectional convolutional operation in the framework, which only passes small-scale information from features of high resolution omitting the propagation of large-scale information. For convenient training, we use ResNet50 as the backbone of all the models, and implement residual attention module [39] instead of SPM. Furthermore, the convolutional operations in UCO and SRM are implemented on 1/4 resolution of input images with 32 channels, and all the models are trained on only official datasets.

As shown in Table IV, compared with the baseline model (UCO), SRM boosts the performance by a large margin in F-measure on MSRA-TD500 and ICDAR2015 respectively, which demonstrates that bi-directional propagation of contextual scale information helps the model generate better representation of multi-size text instances. Furthermore, the consistent activation of multi-size texts can bring a significant improvement to our network (16.5% and 15.7% improvement in F-measure on MSRA-TD500 and ICDAR2015 respectively compared with the baseline model (-)). With slight time consumption, the proposed SRM boosts the performance significantly.

Comparisons between FPN and SRM. As FPN [14] is proposed for handling multi-size object detection, we conduct several experiments to compare our SRM with FPN. For the implementation details of FPN, we remove SRM and add our prediction branch following 4 decoding layers (*F0* to *F3* in Fig. 2). We use 4 hyper-parameters to balance the loss in 4 branches in training. In testing stage, we combine the predictions from 4 branches and use NMS to reduce redundant results. We firstly pre-train all the models on a part of SynthText [43] (about 200 k images), and then finetune them on ICDAR2015 dataset. The results are shown in Table V. As features with different scales have advantages in detecting multi-size texts, the implementation of FPN architecture obtains 0.9% and 0.5% improvement in recall and F-measure respectively. Compared with FPN, our SRM achieves much better results and improves the baseline model in recall, precision and F-measure by 1.8%, 0.3% and 1.1% respectively. We think our method has three advantages: 1) we find the performance of 'baseline + FPN' is sensitive to the balance hyper-parameters in different prediction branches, thus the baseline model implemented with SRM which predicts detection results from **Only One** layer can effectively handle this problem and obtain more accurate results. 2) The supplementation of multi-scale information in SRM helps our method obtain more consistent activation of multi-size texts, which essentially improves the quality of feature maps and is the most important reason for the significant improvement in detection performance. 3) Compared with combining predictions from multi-scale features which causes more redundant detection results, our method spends less time in the NMS processing.

TABLE VI
COMPARISONS WITH DIFFERENT CHANNEL SETTING IN SRM. † AND * MEANS EXPERIMENTS ON ICDAR2015 AND ICDAR2013 RESPECTIVELY

Algorithm	Channel	R	P	F	FPS
R-Net†	128	81.8	**88.8**	85.2	**22.0**
R-Net†	256	**82.8**	88.7	**85.6**	21.4
R-Net*	128	**81.5**	88.5	84.9	**12.0**
R-Net*	256	80.4	**91.1**	**85.4**	11.3

TABLE VII
COMPARISONS OF RESULTS IN DIFFERENT SIZES (TOP) AND RATIOS (BOTTOM). M-P IS SHORT FOR MAX POOLING. B-I IS SHORT FOR BILINEAR INTERPOLATION. SMALL MEANS TEXTS WITH AREA LESS THAN 1000. LARGE MEANS TEXTS WITH AREA LARGER THAN 4500. REMAINING TEXTS ARE DISTRIBUTED IN MIDDLE SIZE (MIDDLE)

Methods	Small	Middle	Large
M-p + B-i	81.1	88.7	83.6
Scale transfer layer	82.0	88.8	84.0
Gain	**+0.9**	**+0.1**	**+0.4**
Methods	$[\frac{1}{2}, 2]$	$[\frac{1}{4}, \frac{1}{2}]$&[2, 4]	else
M-p + B-i	83.0	90.3	81.4
Scale transfer layer	83.1	91.0	82.4
Gain	**+0.1**	**+0.7**	**+1.0**

The ablation study about the number of channels in SRM. We conduct experiments on ICDAR2015 and ICDAR2013 to investigate the relationship between the detection performance and the number of channels in SRM. As shown in Table VI, with slight time consumption, more channels implemented in SRM can obviously boost the performance (0.4% and 0.5% in F-measure on ICDAR2015 and ICDAR 2013 respectively). We ascribe this result to the improvement of representing ability and set the channel to 256 in the remaining experiments.

The evaluation of scale transfer layer. We further conduct several experiments to demonstrate the effectiveness of scale transfer layer for handling texts with large-variance scales. We use max-pooling operation and bilinear interpolation to replace scale transfer layer in our implementation. To make an elaborate evaluation, we compare the results in different sizes and ratios respectively. We firstly pre-train all the models on the whole SynthText [43] for 1 epoch, and then finetune them on ICDAR2015 dataset. As shown in the top of Table VII, benefiting from scale transfer layer, our method can retain more geometrical details of small-size texts compared with max-pooling operation, which improves the detection performance in small-size

TABLE VIII
THE EVALUATION OF DIFFERENT BACKBONE. † AND * MEANS EXPERIMENTS ON ICDAR2015 AND ICDAR2013 RESPECTIVELY

Algorithm	R	P	F
R-Net_ResNet50†	**82.9**	85.1	84.0
R-Net_VGG16†	82.8	**88.7**	**85.6**
R-Net_ResNet50*	77.4	90.9	83.6
R-Net_VGG16*	**80.4**	**91.1**	**85.4**

texts by 0.9% in F-measure. For detecting large-size texts, the proposed scale transfer layer causes much less redundant information compared with bilinear interpolation, which also obtains huge improvement in detection results (0.4% improvement in F-measure). Despite we keep the resolution of *F2* in Fig. 2, benefiting from the supplemented information of middle-size texts from other layers in scale transfer layer, our method can also obtain more accurate results in middle-size text detection (0.1% improvement in F-measure). We also compare the detection results in different ratios to demonstrate the effectiveness of scale transfer layer. As shown in the bottom of Table VII, as the value of ratio increases, the improvement goes from 0.1% to 1%. As mentioned above, the proposed scale transfer layer effectively improves the detection performance of large-variance scale texts.

The evaluation of different backbone. Finally, we implement different backbones in our model and compare their performance in Table VIII. As shown in Table VIII, the proposed method implemented with VGG16 obtains better results in our experiment (this phenomenon also exists in some experiments in [50]). Thus we choose VGG16 as backbone to compare our method with recent approaches.

D. Comparison With Existing Methods

For fair comparison, only the results of existing methods obtained in similar situation are used. To fairly compare the efficiency and accuracy with existing methods, we detect the results in single-scale setting (multi-scale testing sacrifices the efficiency to improve detection accuracy). Thus we only compare our method with the single-scale results with similar resolution of inputs from existing methods [12], [16]–[18], [22], [31], [32], [35]. Furthermore, our method predicts results in the convolutional layer with stride 4, thus we think the detection results predicted from the same stage are more convincing for the comparison of both efficiency and accuracy [32]. Due to recognition branch is proved to effectively improve the performance of detection branch in [12], we show the detection only results in [13], [26] without recognition branch. Following [18], we add the network prediction time and post-processing time together as the inference speed of our method.

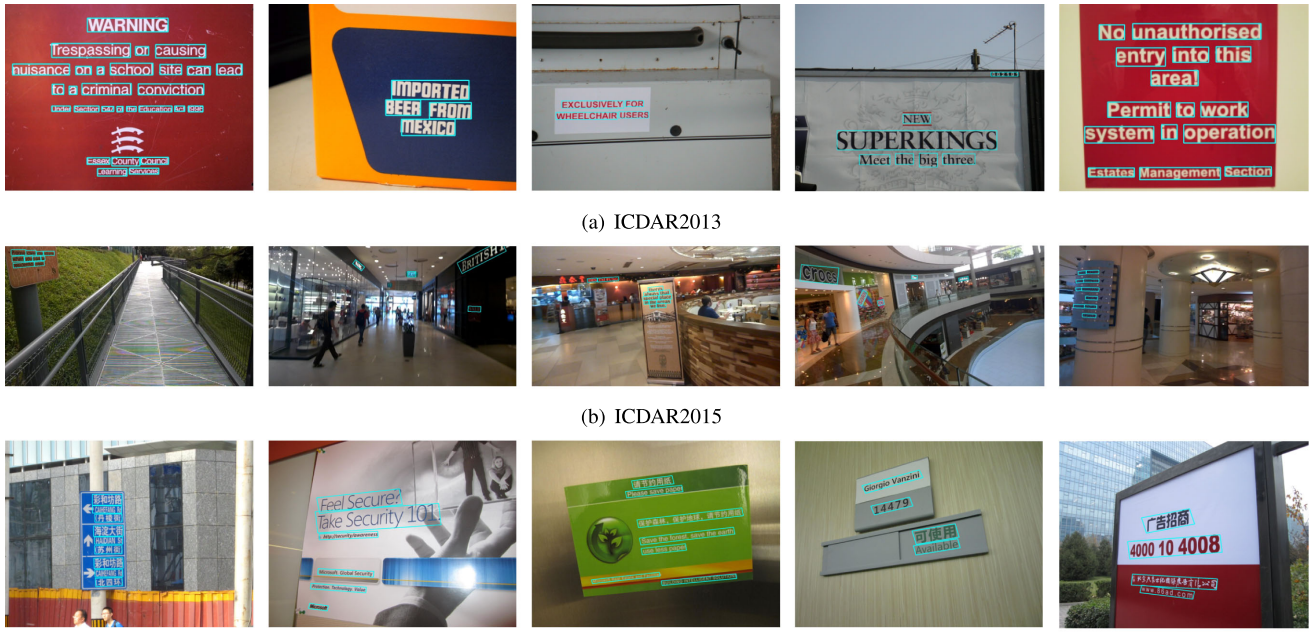
Though the journal version of Mask TextSpotter [53] achieves better performance than the conference version [13], the improvement are mainly exiting in the network implemented with recognition branch, and the results are identical when the network is constructed without recognition branch (e.g. 94.1%, 88.1% and 91.0% in precision, recall and F-measure respectively for det only on ICDAR2013 dataset in both [13], [53]). However, the recognition branch is proved to effectively improve the performance of detection branch in [12], thus, we compare with the performance of 'det only' without recognition for fair comparison which is same to other existing methods [22], [33]. In addition, TIoU [51] which proposes a more advanced metric to quantify the compactness of detection and tightness of matching degree is also used to evaluate the effectiveness of our method.

TABLE IX
RESULTS ON ICDAR2015. TWO-STAGE METHODS ARE ON THE TOP AND ONE-STAGE METHODS ARE ON THE BOTTOM. MULTI-SCALE TESTING AND ENSEMBLE ARE NOT INCLUDED. † MEANS THE BASE NET OF THE MODEL IS NOT VGG16. * MEANS DIRECT REGRESSION BASED METHOD. ATS IS SHORT FOR TRAINED ON ALL THE DATASETS

Algorithm	R	P	F	FPS
RRPN.[2]	73.0	82.0	77.0	4.4
RRPN + ATS.[2]	77.0	84.0	80.0	4.4
TextSpotter[13]†	81.2	85.8	83.4	4.8
FOTS *et al.*[26]†	82.0	88.8	85.3	7.8
Wang *et al.* [25]	**83.3**	**90.4**	**86.8**	**10**
DDR[35]*†	62.0	82.0	70.0	-
EAST[18]*†	73.5	83.6	78.2	13.2
SSTD[11]	73.0	80.0	77.0	7.7
Wang *et al.*[50]*	74.1	85.7	79.5	-
Lyu *et al.*[16]	70.7	**94.1**	80.7	3.6
Textboxes++[12]	76.7	87.2	81.7	11.6
Liao *et al.*[17]	79.0	85.6	82.2	6.5
Pixelink[31]	82.0	85.5	83.7	3
Tian *et al.*[22]†	**84.5**	85.1	84.8	3
PSENet[32]†	83.8	86.1	84.9	3.8
R-Net*	82.8	88.7	**85.6**	**21.4**

1) Evaluation on Oriented Text Benchmark: We evaluate our R-Net on ICDAR2015 dataset to test its performance of arbitrary-orientated text detection. The model is firstly pretrained on SynthText for 1 epoch and then finetuned on ICDAR2015 dataset. We test the results in the original resolution. Quantitative results following the standard evaluation metric is given in Table IX. Since the two-stage methods usually achieve better results in detection task, we list these methods separately on this dataset.

As shown in Table IX, R-Net with single input scale outperforms all existing methods and obtains a new state-of-the-art result (85.6% in F-measure) with the fastest speed (21.4 FPS). As a direct regression based method, R-Net outperforms existing direct regression based methods [18], [35], [50] by a large margin (85.6% vs 78.2%, 70.0% and 79.5%). In addition, our method achieves a much faster speed than existing popular method Textboxes++ [12] (21.4 FPS vs 11.6 FPS). We attribute our high performance to the powerful spatial semantics and consistent activation of multi-size texts generated by SPM and SRM, which essentially improves the quality of feature maps to achieve more accurate detection results. We show the PSENet-4 s of [32] predicted from feature maps with the same resolution as ours (1/4 of input images) for fair comparison in Table IX. Compared with PSENet [32], the proposed method achieves 0.7% improvement in F-measure with much faster speed (21.4 FPS vs 3.8 FPS). Particularly, FOTS [54] without recognition branch is shown here for fair comparison, our R-Net achieves a little better result



(a) ICDAR2013

(b) ICDAR2015

(c) MSRA-TD500

Fig. 8. Results on different datasets. (a) results on ICDAR2013; (b) results on ICDAR2015; and (c) results on MSRA-TD500.

TABLE X
RESULTS EVALUATED WITH TIOU [51] ON ICDAR2015. R_o MEANS RECALL WITH ORIGINAL WORD-LEVEL-ONLY ANNOTATIONS. R_w MEANS RECALL WITH WORD&TEXT-LINE ANNOTATIONS. * INDICATES RESULTS FROM [51]

Algorithm	R_o	P_o	F_o	R_w	P_w	F_w
SegLink [52]*	46.7	58.1	51.7	50.5	59.8	54.8
EAST[18]*	52.8	63.5	57.6	56.7	61.0	60.1
PixelLink[31]*	55.2	61.8	58.3	58.5	62.7	60.5
RRD[17]*	51.5	65.2	57.5	53.0	65.3	58.5
Textbox++ [12]*	53.7	67.2	59.7	59.4	**67.0**	60.3
TextSpotter [13]*	52.7	65.8	58.5	54.9	66.2	60.0
R-Net	**58.7**	**67.9**	**63.0**	**62.3**	66.9	**64.5**

(85.6% vs 85.3% in F-measure) with much faster speed (21.4 FPS vs 7.8 FPS). Furthermore, the proposed method also outperforms proposal-based methods [2], [11]–[13], [16], [17] at least 2.2% in F-measure. Despite RRPN [2] use MSRA-TD500 and ICDAR datasets together to train their model, our R-Net significantly outperforms it by 5.8%, 4.7%, 5.6% in recall, precision and F-measure respectively. Due to the implementation of Squeezeand-Excitation (SE) blocks in backbone in Wang *et al.* [25] (87.6% in F-measure), we only compare our method with [25] implemented without SE blocks for fair comparison. Benefitting from the more concise pipeline with only NMS for post processing, our method can achieve a better trade-off between accuracy and efficiency (10.0 FPS [25] vs our 21.4 FPS).

In addition, we further evaluate our method with a more advanced metric TIoU [51] on ICDAR2015. Results in **Original Word-level-Only Annotations** and **Word&Text-Line Annotations** are shown in Table X. The proposed method achieves the best result in recall precision and F-measure in Original Word-level-Only Annotations and Word&Text-Line Annotations. Thus our method is proved to obtain more tight and accurate detection results compared with existing methods.

Fig. 9. Results on MLT.

2) Evaluation on Long Oriented Text Benchmark: We evaluate our R-Net on MSRA-TD500 dataset to test its ability of long text detection. In test stage, the input images are resized to 768 × 768. Some qualitative results on MSRA-TD500 are depicted in Fig. 8. In general, R-Net achieves satisfactory results in various cases, regardless of fonts, aspect ratios and complex backgrounds.

As shown in Table XI, it is worth mentioning that our R-Net achieves state-of-the-art result with high efficiency. The proposed method significantly surpasses existing direct regression based methods [18], [35], [50] by at least 1.4% in F-measure. Although EAST [18] additionally implements a more powerful PVANET2x as their backbone (achieves 67.4%, 87.3% and

TABLE XI
RESULTS ON MSRA-TD500. * MEANS DIRECT REGRESSION BASED METHOD. EB IS SHORT FOR EMBEDDING. † MEANS THE BASE NET OF THE MODEL IS NOT VGG16

Algorithm	R	P	F	FPS
EAST[18]*	61.6	81.7	70.2	6.5
EAST[18]*†	67.4	87.3	76.1	**13.2**
DDR[35]*†	70.0	77.0	74.0	1.1
Zhang *et al.*[55]	67.0	83.0	74.0	0.48
RRPN [2]	68.0	82.0	74.0	4.4
Xue *et al.**[54]	73.3	80.7	76.8	-
PixelLink[31]	73.2	83.0	77.8	3
TextSnake[30]	73.9	83.2	78.3	1
Liao *et al.*[17]	73.0	87.0	79.0	10.0
Wang *et al.*[50]*†	72.3	**90.3**	80.3	-
Lyu *et al.*[16]	76.2	87.6	81.5	5.7
Tian *et al.*[22]†	76.8	77.2	77.0	-
Tian *et al.* + EB[22]†	**81.7**	84.2	**82.9**	3
R-Net*	79.7	83.7	81.7	11.8

TABLE XII
RESULTS ON ICDAR2013. TWO-STAGE METHODS ARE ON THE TOP AND ONE-STAGE METHODS ARE ON THE BOTTOM. † MEANS THE BASE NET OF THE MODEL IS NOT VGG16. DETEVAL IS USED TO EVALUATE THE RESULTS

Algorithm	R	P	F	FPS
RRPN[2]	72.0	90.0	80.0	4.4
TextSpotter[13]	**88.1**	**94.1**	**91.0**	**4.6**
Textboxes[56]	74.0	88.0	81.0	11.1
Liao *et al.*[17]	75.0	88.0	81.0	-
TextBoxes++[12]	75.0	88.0	81.0	-
Zhang *et al.*[55]	78.0	88.0	83.0	0.5
He *et al.*[57]	76.0	93.0	84.0	4.6
PixelLink[3]	**83.6**	86.4	84.5	3
Lyu *et al.*[16]	79.4	**93.3**	**85.8**	10.4
R-Net	80.4	91.1	85.4	**11.3**

TABLE XIII
RESULTS ON MLT. † MEANS THE BASE NET OF THE MODEL IS NOT VGG16

Algorithm	R	P	F
TH-DL [46]	34.8	67.8	46.0
SARI FDU RRPN V1 [46]	55.5	71.2	62.4
Sensetime OCR [46]	**69.4**	56.9	62.6
SCUT DLVClab1 [46]	54.5	**80.3**	65.0
FOTS [26]†	57.5	79.5	66.7
Lyu *et al.*[16]	55.6	83.8	66.8
R-Net	64.5	70.4	**67.3**

76.1% in recall, precision and F-measure respectively) by doubling the channels of the original backbone, our R-Net also outperforms it by 12.3% and 5.6% in recall and F-measure respectively. Compared with segmentation based methods [30], [31] which contain complex post processing, the proposed method outperforms these methods by a large margin with much faster speed (81.7% vs 78.3% and 77.8% in F-measure and 11.8 FPS vs 3 FPS and 1 FPS). Compared with Tian *et al.* [22], our R-Net outperforms it by 2.9%, 6.5% and 4.7% in recall, precision and F-measure respectively. Although Tian *et al.* [22] achieve better result by implementing an embedding branch, the two-stage clustering algorithm results in large time consumption, and our R-Net can achieve competitive result with much faster speed (11.8 FPS vs 3 FPS). Specifically, Lyu *et al.* [16] combine predictions from seven multi-scale features which have satisfying representation for large-variance scale texts. However, our R-Net can achieve better results (81.7 % vs 81.5 % in F-measure) and much faster speed (5.7 FPS vs 11.8 FPS) by making prediction on only one convolutional layer. Compared with proposal-based methods [2], [16], [17], [54], [55], the proposed method achieves much better performance with faster speed. Further more, proposal-based methods are sensitive to the anchor setting, but our R-Net, which directly regresses the distances to four edges without complex hand-crafted hyperparameter setting, is a more robust approach for scene text detection.

3) Evaluation on Horizontal Text Benchmark: We evaluate our R-Net on ICDAR2013, one of the most popular horizontal text datasets, to test its ability of horizontal text detection. The input images in test stage are resized to 768 $\times$ 768. Since the two-stage methods usually achieve better results in detection task, we list these methods separately in this dataset.

As shown in Table XII with the help of SPM and SRM, the proposed method achieves state-of-the-art results in recall, precision and F-measure respectively, outperforming most existing methods (e.g., RRPN [2], Liao*et al.* [17], TextBoxes++ [12]) by a large margin with much faster speed (11.3 FPS). Though Textboxes [56] achieves a comparable speed to ours (11.1 FPS vs 11.3 FPS), R-Net outperforms it by 6.4%, 3.1% and 4.4% in recall, precision and F-measure respectively. Segmentation based method PixelLink [31] predicts text regions by separating pixels into different text instances according to the pixel linkages. However, this pixel-wise operation results in large time consumption. Compared with PixelLink [31], our R-Net achieves 4.7 % and 0.9 % improvement in precision and F-measure respectively with much faster speed (11.3 FPS vs 3 FPS). Though the proposal-based method [16] obtains a little better results than ours in this dataset, the proposed method can achieve a better trade-off between accuracy and efficiency with a more concise pipeline. Besides, our R-Net can run at 11.3 FPS, which is much faster than other existing methods.

4) Evaluation on Multi-Lingual Benchmark: To verify the ability of R-Net to detect multi-lingual texts, we evaluate R-Net on MLT dataset by finetuning our model about 110 epochs from SynthText. The results are evaluated online and compared with existing state-of-the-art methods in Table XIII. When testing at single scale, R-Net can achieve 64.5%, 70.4% and 67.3% in recall, precision and F-measure respectively, which outperforms existing popular methods (FOTS [26] and Lyu *et al.* [16]). As only a few listed methods publish their speed, we only show the accuracy on this dataset for fair comparison.

5) Evaluation on Curved Text Benchmark: Due to the limitation of the representation of texts used in our approach (we predict the offsets form point to corresponding 4 edges), our method is proposed for detecting quadrangular texts. However, we think it is necessary to compare our method with parallel quadrangular-text detection methods on curved text benchmark. As shown in Table XIV, we conduct additional experiments on CTW1500 [58]. We simply generate the minimum enclosing

TABLE XIV
RESULTS ON CTW1500. * INDICATES RESULTS IN [58]

Algorithm	R	P	F
CPTN [59]*	53.8	60.4	56.9
SegLink [52]*	40.0	42.3	40.8
EAST [18]*	49.1	**78.7**	60.4
R-Net	**71.0**	74.6	**72.8**

TABLE XV
COMPARISONS OF BOTH TRAINING TIME AND TESTING TIME. F IS SHORT FOR F-MEASURE. TR MEANS TRAINING TIME (HOURS). TE MEANS TESTING TIME(FPS). * AND † MEANS RESULTS ON ICDAR2015 AND MSAR-TD500 RESPECTIVELY.

Methods	F*	Tr*	Te*	F †	Tr †	Te †
PixelLink [31]	83.7	**31**	3	77.8	**13**	3
TextSnake [30]	82.6	36	1.1	78.3	15	1.1
R-Net	**85.6**	54	**21.4**	**81.7**	35	**11.8**

TABLE XVI
THE PERFORMANCE GAIN OF PROPOSED SPM AND SRM IN A STRONGER BASELINE

SPM	R	P	F
Mask-RCNN	79.5	85.5	82.4
Mask-RCNN+SPM	80.8	86.7	83.6
Mask-RCNN+SRM	81.1	86.8	83.8
Mask-RCNN+SPM+SRM	**81.4**	**88.3**	**84.7**

TABLE XVII
THE PERFORMANCE GAIN OF PROPOSED SPM AND SRM IN ANOTHER BASELINE MODEL

SPM	R	P	F
PSENet	74.7	80.1	77.3
PSENet+SRM	75.0	81.8	78.2
PSENet+SPM+SRM	**75.2**	**83.5**	**79.1**

rectangles from official polygon annotations in training stage as bounding box labels. Compared with parallel quadrangular-text detection methods, our R-Net achieves the best results and outperforms these method by at least 12.4% in F-measure, which demonstrate the effectiveness of our method. Compared with the same DRBM [18], the proposed method outperforms it by a large margin (72.8% vs 60.4% in F-measure).

E. Comparisons of Both Training Time and Testing Time

We compare both training time and testing time on ICDAR2015 and MSRA-TD500 with exiting methods in Table XV. As recent methods usually do not show the training time in their papers, we select two representative methods (PixelLink [31] and TextSnake [30]) and reproduce their released code in our environment. To be specific, all the models are trained on 1 NVIDIA TITANX Pascal GPU with batch size of 8 for fair comparison. We refer to the training details from their paper and convert the corresponding training steps into our settings. For the testing time, we simply refer to their paper. Our method spends more time in the back-propagation computation in training stage, however, the concise pipeline in our method with only NMS for post-processing help our method obtain higher efficiency in testing stage compared with existing methods.As shown in Table XV (our method: batch size of 8 on 1 GPU for 75 k iterations on ICDAR2015 dataset and batch size of 8 on 1 GPU for 52 k iterations on MSRA-TD500 dataset), our method achieves comparable training time compared with existing methods and outperforms them by a large margin in both accuracy and efficiency.

F. Rationality of High Performance and Fast Speed

R-Net is proposed to detect multi-size texts with arbitrary direction. The huge improvement in accuracy and efficiency is mainly due to three points: 1) we design a Spatial Relationship Module to model the spatial dependencies and independencies simultaneously. SPM enhances the long-range dependence information between pixels at different locations and also boosts the discriminative features in local visual cues. 2) We propose an effective Scale Relationship Module to essentially enhance the representing ability of large-variance scale texts. The consistent activated features reduce the change in the scale space, and help the model obtian more accurate results. 3) The light network and the robust regression approach are the keys to make our model efficient and accurate.

To further evaluate the effectiveness of SRM and SPM, we embed proposed modules into a stronger baseline (2-stage methods) Mask-RCNN step by step. To be specific, we implement SPM in the last encoding layer in backbone and embed SRM to the decoding layers in backbone. For training stage, we only use the official images in ICDAR2015 to train all the models (without pre-training on SynthText [43]). Data augmentation includes random rotation, random horizontal flip and random crop. As shown in Table XVI, when we implement both proposed modules in Mask-RCNN, the improvement increase to 2.3% in F-measure. Without any other modifications, our SPM and SRM significantly improve the detection performance of Mask-RCNN and help it achieve the state-of-the-art results, which suppress most of the recent methods in Table IX without external training images.

Finally, we further embed the proposed modules in PSENet [32] (PSE-4 s is used in this experiment) to demonstrate the effectiveness of our method. We first pre-train the model on MLT dataset [46] for about 70 epoches and then finetune it on CTW dataset [58] for 384 epoches. As shown in Table XVII, the proposed modules effectively improve the performance of detection results. When we embed both SPM and SRM in the baseline model, the improvement increases to 1.8%. In addition, we infer more training steps used in the pre-training stage can further help the proposed modules to learn samples and improve the results.

G. Limitations

As demonstrated by preceding experimental results, the proposed R-Net can perform well in most situations, effectively handling large-variance scale problem and accurately detecting texts in complex background. However, it fails to handle texts

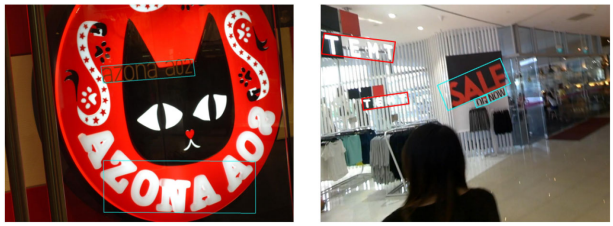


Fig. 10. Red boxes are ground truths while blue boxes are predicted results.

with large character spacing (Fig. 10). Due to the limitation of quadrilateral annotations, R-Net also performs not well in curved text detection. Note that these difficulties are also existing in other state-of-the-art methods [2], [12], [16].

V. Conclusion

In this paper, we propose a new end-to-end method for efficient and accurate scene text detection called R-Net. R-Net aims to essentially address the large-variance scale problem by mapping the multi-scale features to a scale-invariant space through Scale Relationship Module (SRM). We also visualize the feature maps to prove the effectiveness of our approach. To further enhance the representation ability of scene texts in complex background, a novel Spatial Relationship Module (SPM) is proposed to enhance the spatial semantics by considering the long-range dependencies and local independencies simultaneously. In all experiments, the proposed method achieves state-of-the-art performance with high efficiency for horizontal, multi-oriented, long text and multi-lingual cases. The exhaustive ablation studies demonstrate the internal relations of each setting and can generalize to other works to give insights on some important problems. In the feature, we will combine our method with a recognition branch to further improve the performance.

References

[1] X. Ren *et al.*, "A convolutional neural network-based chinese text detection algorithm via text structure modeling," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 506–518, Mar. 2017.

[2] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.

[3] L. Wu, P. Shivakumara, T. Lu, and C. L. Tan, "A new technique for multi-oriented scene text line detection and tracking in video," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1137–1152, Aug. 2015.

[4] X. Liu and W. Wang, "Robustly extracting captions in videos based on stroke-like edges and spatio-temporal analysis," *IEEE Trans. Multimedia*, vol. 14, no. 2, pp. 482–489, Apr. 2011.

[5] X.-C. Yin, Z.-Y. Zuo, S. Tian, and C.-L. Liu, "Text detection, tracking and recognition in video: A comprehensive survey," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2752–2773, Jun. 2016.

[6] S. Karaoglu, R. Tao, T. Gevers, and A. W. Smeulders, "Words matter: Scene text for image classification and retrieval," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1063–1076, May 2017.

[7] H. Xie *et al.*, "Convolutional attention networks for scene text recognition," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 1s, pp. 1–17, 2019.

[8] J. Li *et al.*, "Attentive contexts for object detection," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 944–954, May 2017.

[9] C. Chen and Q. Ling, "Adaptive convolution for object detection," *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 3205–3217, Dec. 2019.

[10] K. L. Bouman, G. Abdollahian, M. Boutin, and E. J. Delp, "A low complexity sign detection and text localization method for mobile applications," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 922–934, Oct. 2011.

[11] P. He *et al.*, "Single shot text detector with regional attention," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 3047–3055.

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

[13] 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 *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 67–83.

[14] T. Lin *et al.*, "Feature pyramid networks for object detection," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 936–944.

[15] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 21–37.

[16] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7553–7563.

[17] M. Liao, Z. Zhu, B. Shi, G. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 5909–5918.

[18] X. Zhou *et al.*, "EAST: An efficient and accurate scene text detector," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2642–2651.

[19] Y. Wang, H. Xie, Z. Fu, and Y. Zhang, "DSRN: A deep scale relationship network for scene text detection," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 947–953.

[20] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 102–114, Jan. 2017.

[21] A. Liu *et al.*, "Multi-level policy and reward reinforcement learning for image captioning." in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 821–827.

[22] Z. Tian *et al.*, "Learning shape-aware embedding for scene text detection," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 4234–4243.

[23] Y. Pang, T. Wang, R. M. Anwer, F. S. Khan, and L. Shao, "Efficient featurized image pyramid network for single shot detector," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 7336–7344.

[24] Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Oriented response networks," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 4961–4970.

[25] X. Wang *et al.*, "Arbitrary shape scene text detection with adaptive text region representation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 6449–6458.

[26] X. Liu *et al.*, "FOTS: Fast oriented text spotting with a unified network," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 5676–5685.

[27] Z. Liu *et al.*, "Towards robust curve text detection with conditional spatial expansion," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 7269–7278.

[28] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3431–3440.

[29] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 3146–3154.

[30] S. Long *et al.*, "TextSnake: A flexible representation for detecting text of arbitrary shapes," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 19–35.

[31] D. Deng, H. Liu, X. Li, and D. Cai, "PixelLink: Detecting scene text via instance segmentation," in *Proc. Assoc. Advancement Artif. Intell.*, 2018, pp. 6773–6780.

[32] W. Wang *et al.*, "Shape robust text detection with progressive scale expansion network," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Jun. 2019, pp. 9328–9337.

[33] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 9365–9374.

[34] Z. Liu *et al.*, "Learning Markov clustering networks for scene text detection," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 6936–6944.

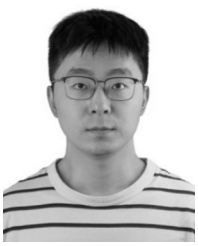
[35] W. He, X. Zhang, F. Yin, and C. Liu, "Deep direct regression for multi-oriented scene text detection," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 745–753.

[36] X. Zeng, W. Ouyang, B. Yang, J. Yan, and X. Wang, "Gated bi-directional CNN for object detection," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 354–369.

[37] T. He *et al.*, "An end-to-end textspotter with explicit alignment and attention," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 5020–5029.

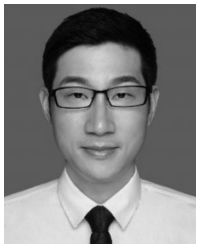
[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.

[39] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 3156–3164.
[40] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7794–7803.
[41] P. Zhou, B. Ni, C. Geng, J. Hu, and Y. Xu, "Scale-transferrable object detection," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 528–537.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.
[43] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2315–2324.
[44] D. Karatzas *et al.*, "ICDAR 2013 robust reading competition," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, 2013, pp. 1484–1493.
[45] D. Karatzas *et al.*, "ICDAR 2015 competition on robust reading," in *Proc. 13th Int. Conf. Document Anal. Recognit.*, 2015, pp. 1156–1160.
[46] N. Nayef *et al.* "Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-RRC-MLT," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit.*, vol 1, 2017, pp. 1454–1459.
[47] C. Yao, X. Bai, and W. Liu, "A unified framework for multioriented text detection and recognition," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4737–4749, Nov. 2014.
[48] B. Shi *et al.*, "ICDAR2017 competition on reading chinese text in the wild (RCTW-17)," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit.*, vol. 1, 2017, pp. 1429–1434.
[49] H. Tang *et al.*, "Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 2417–2426.
[50] F. Wang, L. Zhao, X. Li, X. Wang, and D. Tao, "Geometry-aware scene text detection with instance transformation network," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 1381–1389.
[51] Y. Liu *et al.*, "Tightness-aware evaluation protocol for scene text detection," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 9612–9620.
[52] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2550–2558.
[53] M. Liao *et al.*, "Mask textSpotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019, p. 1.
[54] C. Xue, S. Lu, and F. Zhan, "Accurate scene text detection through border semantics awareness and bootstrapping," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 355–372.
[55] Z. Zhang *et al.*, "Multi-oriented text detection with fully convolutional networks," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4159–4167.
[56] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: A fast text detector with a single deep neural network," in *Proc. Assoc. Advancement Artif. Intell.*, 2017, pp. 4161–4167.
[57] T. He, W. Huang, Y. Qiao, and J. Yao, "Text-attentional convolutional neural network for scene text detection," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2529–2541, Jun. 2016.
[58] Y. Liu, L. Jin, S. Zhang, C. Luo, and S. Zhang, "Curved scene text detection via transverse and longitudinal sequence connection," *Pattern Recognit.*, vol. 90, pp. 337–345, 2019.
[59] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 56–72.



Yuxin Wang received the B.S. degree from XiDian University in 2018, he is currently working toward the M.S. degree with the School of Information Science and Technology, University of Science and Technology of China. His research interests mainly cover computer vision and signal processing.

Hongtao Xie received the Ph.D. degree in computer application technology from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2012. He is currently a Research Professor with the School of Information Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include multimedia content analysis and retrieval, deep learning and computer vision.



Zhengjun Zha received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2004 and 2009, respectively. He is currently a Full Professor with the School of Information Science and Technology, University of Science and Technology of China, the Vice Director of National Engineering Laboratory for Brain-Inspired Intelligence Technology and Application. He was a Researcher with the Hefei Institutes of Physical Science, Chinese Academy of Sciences, from 2013 to 2015, a Senior Research Fellow with the School of Computing, National University of Singapore (NUS), from 2011 to 2013, and a Research Fellow there from 2009 to 2010. He has authored or coauthored more than 100 papers in these areas with a series of publications on top journals and conferences. His research interests include multimedia analysis, retrieval and applications, as well as computer vision etc. Prof. Zha was the recipient of multiple paper awards from prestigious multimedia conferences, including the Best Paper Award and Best Student Paper Award in ACM Multimedia, etc.

Youliang Tian received the B.Sc. degree in mathematics and applied mathematics in 2004, the M.Sc. degree in applied mathematics from GuiZhou University in 2009, and the Ph.D. degree in cryptography from Xidian University in 2012. He is now a professor and Ph.D. supervisor with College Of Computer Science & Technology, GuiZhou University. In the years 2012 to 2015 he was a Postdoctoral Associate at the State Key Laboratory for Chinese Academy of Sciences. His research interests include algorithm game theory, cryptography and information security protocol, etc. He is the academic leader of Big Data Privacy Protection and Data Security in The State Key Laboratory of Public Big Data of Guizhou Province, the vice chairman of Southwest BBS of Chinese Association for Artificial Intelligence, the deputy director of the Institute of Cryptography & Data Security and the deputy dean of the department of Cyberspace Security in Guizhou university, etc. He is also the editorial board member of Journal on Communications and Chinese Journal of Network and Information Security.

Zilong Fu received the B.S. degree from XiDian University in 2018 and He is currently working toward the M.S. degree with the School of Information Science and Technology, University of Science and Technology of China. His research interests mainly cover computer vision and signal processing.

Yongdong Zhang (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from Tianjin University, Tianjin, China, in 2002. He is currently a Professor with the School of Information Science and Technology, University of Science and Technology of China, Hefei, China. He has authored more than 100 refereed journal and conference papers. His research interests include multimedia content analysis and understanding, multimedia content security, video encoding, and streaming media technology. Prof. Zhang was the recipient of the Best Paper Awards in PCM 2013, ICIMCS 2013, and ICME 2010, and the Best Paper Candidate in ICME 2011. He serves as an Editorial Board Member of the Multimedia Systems Journal and the IEEE TRANSACTIONS ON MULTIMEDIA.