ADNet: Rethinking the Shrunk Polygon-Based Approach in Scene Text Detection

Yadong Qu, Hongtao Xie*, Shancheng Fang, Yuxin Wang, Yongdong Zhang, *Senior Member, IEEE*

***Abstract*—To localize text regions and separate close instances, the shrunk polygon is widely used in recent scene text detection methods. However, there exist two problems: 1) Existing methods fail to consider the aspect ratio sensitive problem when reconstructing the text instance from shrunk polygon. 2) Texts with extreme aspect ratios will lead to the fracture of shrunk polygons. To handle these two problems, in this paper, we propose a novel Adaptive Dilation Network (ADNet) to focus on the reconstruction process from shrunk polygon, which aims to provide a tight and complete text representation. Firstly, instead of using a fixed dilation factor, ADNet uses an aspect ratio-wise dilation factor to reconstruct the text region from shrunk polygon for each text instance. Such an instance-wise dilation factor considers the scale correlation between the original and shrunk polygon, and thus can guide an adaptive text region reconstruction for texts with large aspect ratio variance. Secondly, to deal with the fracture of detection results, a new Efficient Spatial Relationship Module (ESRM) is devised to capture long-range dependencies with low computation cost. ESRM uses a novel Weighted Pooling to reduce the resolution of feature maps without much information loss. Compared with the existing methods, ADNet further explores the potential of shrunk polygon-based approaches and obtains excellent detection results at an impressive speed. Extensive experiments on several datasets (Total-Text, CTW1500, MSRA-TD500 and ICDAR2015) verify the superiority of our method. Code will be available at https://github.com/qqqyd/ADNet.**

***Index Terms*—Scene text detection, shrunk polygon, aspect ratio, adaptive dilation factor.**

I. Introduction

THE scene text detection task aims to locate text areas in the complex background. It is a fundamental step for end-to-end text spotting, which has a wide range of applications such as intelligent transportation systems and wearable translation equipment. The segmentation-based methods [1], [2], [3], [4] become popular for their capacity of detecting arbitrary-shaped text instances. Benefiting from the impressive performance in shape representation and separation of close text instances, shrunk polygon plays a dominant role in segmentation-based methods. Although many works have achieved outstanding performance by applying shrunk polygon, there are still two problems that remain unsolved.

The first problem is how to accurately rebuild the text region from the predicted shrunk polygon. PSENet [2] predicts a series of kernel regions with different sizes and gradually integrates the smallest kernel into a large one. However, it is difficult to balance the learning process between the kernel maps with different sizes. Without the pixel-level reconstruction process, DB [1] generates a single kernel map and empirically presets a dilation factor R_d to construct all the kernels. DB [1] proves that a single kernel map is enough to represent the shape and position of a text instance, and the complicated post-processing is also avoidable. However, as shown at the top of Fig. 1, although the shrunk polygon is well perceived, assigning a constant dilation factor for all reconstructions is still suboptimal and leads to coarse boundary predictions (red polygons). The incomplete detection results will make the subsequent text recognition process [5], [6], [7], [8] unable to fully perceive the character features, thus reducing the recognition performance. We attribute such an inaccurate detection to the limitation of the fixed dilation factor for texts with variable aspect ratios (demonstrated formally in Sec.III.A). It is worth mentioning that we are the first one to point out this problem in the scene text detection

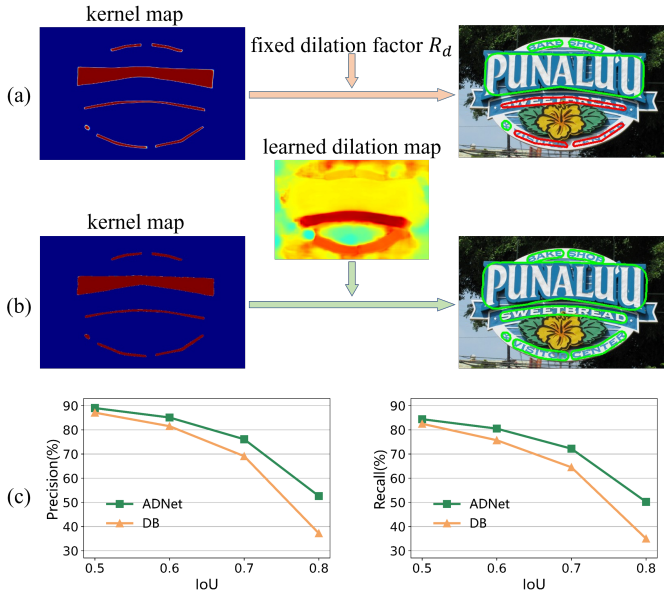


Fig. 1. (a) and (b) are the visualizations of DB [1] and our ADNet. The fine and rough boundary descriptions are shown in green polygons and red polygons, respectively. (c) shows the results of precision and recall of DB and ADNet under different evaluation metrics (IoU@0.5, 0.6, 0.7, 0.8). Kernel maps of the two methods are almost the same. Benefiting from the assignment of independent dilation factor to each polygon through the predicted dilation map, our ADNet significantly represents the arbitrary-shaped texts with more accurate and tight boundaries.

This work is supported by the National Nature Science Foundation of China (62121002, 62022076, U1936210), the Youth Innovation Promotion Association Chinese Academy of Sciences (Y2021122).

Y. Qu, H. Xie, S. Fang, Y. Wang, Y. Zhang are with School of Information Science and Technology, University of Science and Technology of China (e-mail: qqqyd@mail.ustc.edu.cn; htxie@ustc.edu.cn; fangsc@ustc.edu.cn; wangyx58@mail.ustc.edu.cn; zhyd73@ustc.edu.cn).

H. Xie is the corresponding author.

and summarize this issue as the *Shrunk polygon Aspect ratio Sensitive* (SAS) problem. Thus, it is necessary to explore a new effective reconstruction method based on the shrunk polygon.

The second problem is the fracture of text detection results with extreme aspect ratios. Due to the limited receptive field of CNNs, it is difficult for general convolution kernels to capture the complete features of long texts, resulting in detecting text instances into multiple small ones. Although simply stacking the convolution layers can effectively alleviate such a problem, the complexity of the network often makes it hard to optimize. It is also unrealistic to design a convolution kernel that meets all aspect ratios. Even though the existing attention mechanism [9], [10], [11], [12] is instructive to capture the context information and detect long texts, these methods mainly focus on capturing global dependencies while ignoring the amount of calculation. The enormous extra computation cost limits the efficiency of these approaches.

In this paper, we introduce an Adaptive Dilation Network (ADNet), including a novel Adaptive Dilation Module (ADM) and an Efficient Spatial Relationship Module (ESRM) to solve the above two problems, respectively. As shown in Fig. 2, the ADM is designed following two characteristics: (1) ADM adopts a Relationship Linking Module to reason the inner connection between the original and shrunk text region. Such correlation information helps to infer a more accurate instance-wise dilation factor for the reconstruction of each text. (2) To ensure a detailed boundary description, ADM is supervised under an aspect ratio-aware dilation map. The dilation map is generated by taking the aspect ratio of the text region into account. Furthermore, to improve the performance on large-aspect-ratio texts, we propose the ESRM to enhance the feature representation ability by capturing long-range dependencies in visual contexts. Compared with previous long-range relationship capturing modules [9], [10], [11], the proposed ESRM introduces a novel Weighted Pooling (Wei-Pooling) to down-sample the feature map while preserving the information as much as possible. In the following Spatial Relationship Module, the spatial dependencies can be calculated under a small resolution, which guarantees a high efficiency without much information loss. Extensive experiments demonstrate that the proposed ADNet achieves state-of-the-art performance on several public benchmarks (Total-Text, CTW1500, MSRA-TD500 and ICDAR2015) containing long and arbitrary shape texts with an impressive speed.

The main contributions of our work are three-fold:

- We are the first to point out the limitation of fixed dilation factor for variable aspect ratio text reconstruction, which is called Shrunk polygon Aspect ratio Sensitive problem.
- We propose a novel Adaptive Dilation Module to predict an independent aspect ratio-aware dilation factor for each text instance. Benefiting from exploiting the relationship between shrunk and original text regions, ADM obtains a more accurate and tight boundary representation for arbitrary-shaped texts.
- The Efficient Spatial Relationship Module is proposed to capture long-range dependencies with little computation cost, which uses a new Weighted Pooling operation to lighten the computation burden without losing much context information.

II. Related Work

Scene text detection has been greatly improved since deep learning has become popular [13], [14]. Existing methods can be roughly divided into two categories: regression-based methods and segmentation-based methods. At the same time, as one of the segmentation-based methods, the shrunk polygon-based method occupies an increasingly important position due to its excellent performance in separating adjacent texts.

A. Regression-based Methods

Most regression-based methods are inspired by the common object detectors [15], [16], [17]. This kind of method usually generates text bounding boxes by predicting the offsets from preset anchors or pixels to the corresponding ground truth. CTPN [18] detects a series of small components by regressing the position of anchors with a constant width and then connects them into a whole. Liao *et al.* [19] first propose Textboxes, which adopts the irregular convolution filters to handle text instances with large aspect ratios. Then they make improvements and propose Textboxes++ [20], which adds an angle prediction to detect multi-oriented texts. Instead of regressing the vertices of the bounding box, DDR [21] and EAST [22] directly predict the offsets from vertices or boundaries to the current pixel. Although these methods achieve good performance for horizontal or multi-oriented texts, they fail to describe text with irregular shapes, such as curved texts.

To detect curve texts, based on EAST [22], LOMO [23] proposes a shape expression module to predict the offsets from pixels in a text to the related border points. Moreover, LOMO [23] performs iterative optimization to refine the regressed bounding boxes and gradually perceive the whole text instances. Other than regressing boundary points in the Cartesian system, TextRay [24] detects arbitrary-shaped texts in the polar system, which regresses the distance between the boundary points and the text center. However, the regression error mentioned in [25] cannot be ignored. In addition, there is still much room to improve the tightness of the detection results for curved texts.

B. Segmentation-based Methods

Segmentation-based methods mainly use FCN [26] to obtain a pixel-level prediction to reconstruct the text region. A majority of these methods follow the bottom-up paradigm, which means first obtain the expected text components and then combine them according to a preset rule. PixelLink [3] groups the pixels belonging to the same text based on the linkage relationship with neighbor pixels. TextField [4] and TextMountain [27] both introduce a direction map indicating the direction from a pixel to the corresponding nearest boundary to separate adjacent texts. Tian *et al.* [28] adopts a clustering strategy by mapping the pixels to an embedding space to distinguish each text instance.

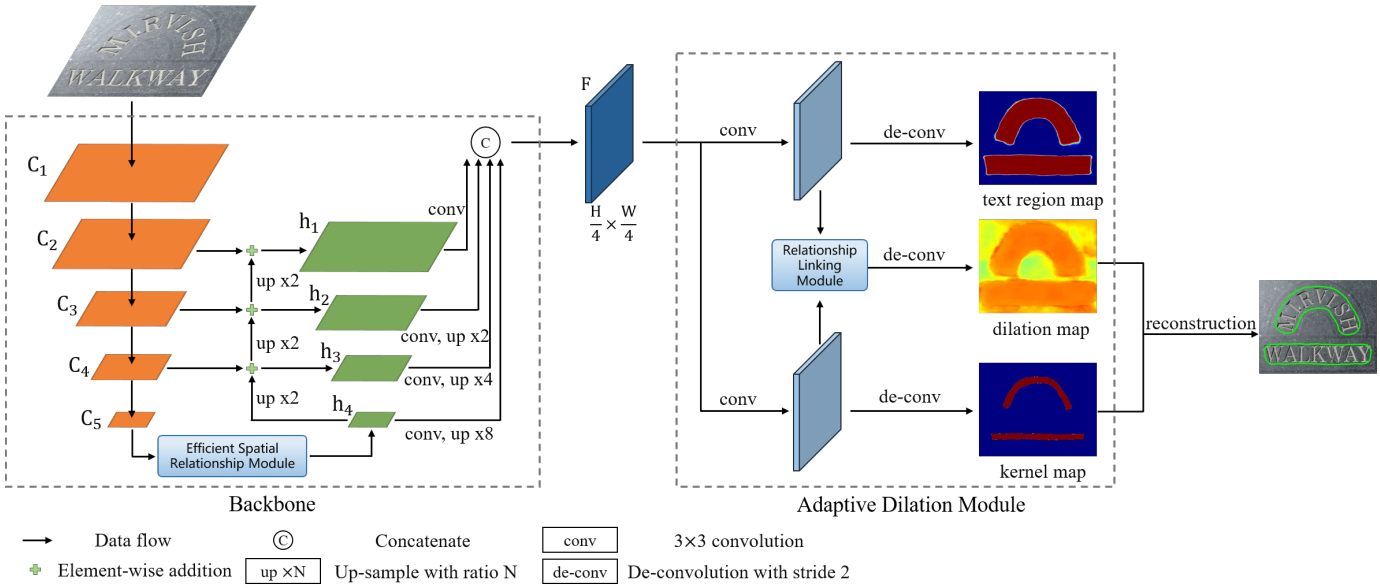


Fig. 2. The overall pipeline of ADNet. The network consists of three parts: backbone, Efficient Spatial Relationship Module (ESRM) and Adaptive Dilation Module (ADM). The backbone is adapted from ResNet50 and applied with ESRM. ADM takes the fused feature as input and outputs three maps. Only dilation map and kernel map are needed for reconstruction.

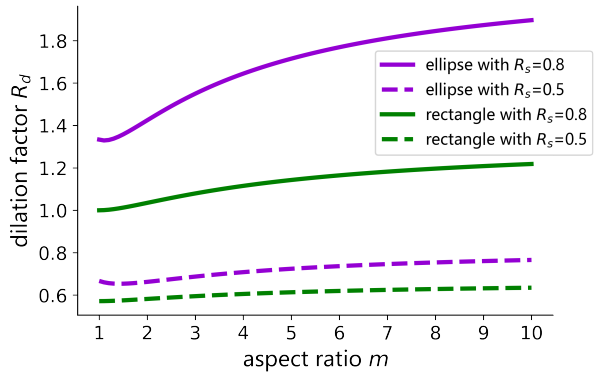


Fig. 3. Illustration of Eq. (3) and Eq. (4). The purple and green lines represent the cases modeled as ellipses and rectangles, respectively.

As segmentation-based methods can obtain pixel-level results, which provide the flexibility to aggregate the components, they have great advantages in detecting curved texts. Among them, due to the excellent performance of shrunk polygon in separating the adjacent texts, the shrunk polygon is adopted in more and more methods [1], [2], [25], [29], [27], [30]. PSENet [2] uses a series of shrunk polygons with different sizes to represent a text region. The smallest polygon is used to distinguish different texts, but it lacks the information of the text boundary. Progressively integrating the small kernel into a large one can generate an accurate detection result. DB [1] directly predicts a shrunk region map and expands it with a constant dilation factor to reduce the complexity of post-processing and improve the speed. The text center line generated in TextSnake [29] and TextDragon [31] is also a kind of shrunk polygon. It is not used for direct expansion to obtain the bounding box, but as auxiliary information to assist the reconstruction process.

Nevertheless, the complicated expansion algorithm in [2] severely limits the speed. The bounding boxes generated from [1] fail to encircle the text area tightly, which will diminish the performance of subsequent text recognition in the text spotting task [32], [33]. Different from the above methods, our ADNet proposes a simple and effective module to reconstruct the text region by adaptively dilating the shrunk region. Especially, ADNet is the first work to explore the Shrunk polygon Aspect ratio Sensitive (SAS) problem.

III. Methodology

A. Preliminary Knowledge

Existing methods [1], [2], [25], [30] obtain the kernel region label by shrinking the polygon G with distance D using the Vatti clipping algorithm [34]. The distance D is calculated as follows:

$$D = \frac{R_s \cdot S_g}{C_g}, \tag{1}$$

where S_g and C_g are the area and perimeter of G. R_s is the preset shrink factor. Similarly, Eq. (1) is used when calculating the dilation distance that a shrunk polygon needs to be constructed to the original polygon after getting the predicted shrunk region map. The only differences are that R_s is replaced by a preset fixed dilation factor R_d, and (S_g, C_g) are replaced by the area and perimeter (S_k, C_k) of the shrunk polygon. However, the fixed R_d cannot adapt well to various text regions due to a large variety of aspect ratios.

SAS problem. Without losing generality, we model the text instance as an ellipse with the major axis of a and minor axis of b. To rebuild the original area, the shrinking distance and dilation distance should be consistent, which means

$$\frac{R_d \cdot S_k}{C_k} = \frac{R_s \cdot S_g}{C_g}. \tag{2}$$

By simplifying Eq. (2) we can obtain:

$$R_d = \frac{pR_s(p - R_s) + (2 - \frac{4}{\pi})(1 - m)R_s}{(p - R_s)(p - mR_s)}, \tag{3}$$

where $m = \frac{a}{b}$, indicating the aspect ratio of the original text instance, and $p = 2 - \frac{4}{\pi} + \frac{4}{\pi}m$. The detailed derivation process is shown in the Appendix.

Similarly, if we model the text instance as a rectangle with width and height equal to a and b, Eq. (2) can be simplified to:

$$R_d = \frac{-4R_s + 16}{\frac{m}{(m+1)^2}R_s^2 - 2R_s + 4} - 4, \tag{4}$$

where $m = \frac{a}{b}$.

Eq. (3) and Eq. (4) can be unified as $R_d = f(R_s, m)$, which indicates that R_d is not only related to R_s, but also to the aspect ratio m. In other words, a fixed R_d set for an invariant R_s can only be applied to a certain aspect ratio. Fig. 3 shows the variation of R_d with m when R_s is fixed. We can see that whether we model the text instance as a rectangle or an ellipse, a larger aspect ratio corresponds to a larger R_d. Thus, as shown in Fig. 1, texts with a large aspect ratio variance cannot be reconstructed accurately by applying a constant R_d for all shrunk polygons (red polygons), which is called Shrunk polygon Aspect ratio Sensitive (SAS) problem. Therefore, we design a novel Adaptive Dilation Module to learn an independent aspect ratio-aware dilation factor for each text instance and accurately rebuild the whole text region by exploiting the relationship between the original and shrunk text region features.

B. Pipeline

As illustrated in Fig. 2, the overall pipeline of the proposed network consists of three parts: backbone, Efficient Spatial Relationship Module (ESRM) and Adaptive Dilation Module (ADM). Firstly, ResNet50 [35] is adopted as our basic network to extract the pyramid feature of an input image. Then, the feature from the last stage of ResNet50 is fed into ESRM to enhance the global contextual semantic representation with little computation cost. To aggregate the multi-scale information, inspired by FPN [36], we merge feature maps from different stages through element-wise addition and concatenate the multi-stage features to generate the final feature map F. Next, ADM takes F as input and outputs text region map, kernel map and aspect ratio-aware dilation map. The text region map is used to help the kernel branch perceive the original shape information of the text instance. In the inference stage, the bounding box can be easily generated from the kernel map and aspect ratio-aware dilation map.

The details about ESRM and ADM are illustrated in Sec.III.C and Sec.III.D, respectively.

C. Efficient Spatial Relationship Module

Due to the lack of global contextual information, the general scene text detection model usually fails to handle the texts with extreme aspect ratios, leading to the fracture in detection results. Although using the attention mechanism [9], [10], [11], [12] to capture long-range dependencies is informative to solve this problem, the extra computation cost heavily limits the efficiency of these methods. To alleviate this problem, an Efficient Spatial Relationship Module (ESRM) is designed to capture global contextual information while maintaining high efficiency.

The architecture of ESRM is shown in Fig. 4. It mainly contains two parts: Weighted Pooling (Wei-Pooling) and Spatial Relationship Module (SRM). Wei-Pooling is proposed to reduce the resolution of the feature map with little loss of information. Particularly, it firstly generates a weighted map $T \in \mathbb{R}^{1\times H\times W}$ from a 3×3 convolution layer. Then coefficients in each grid with size $N \times N$ are normalized by the softmax function. Finally, the normalized weighted map and feature map C_5 are multiplied and summed in each grid to obtain the pooled feature map $C' \in \mathbb{R}^{C\times \frac{H}{N}\times \frac{W}{N}}$. The output feature map y of the Wei-Pooling at location p is formulated by Eq. (5).

$$y_p = \frac{1}{N \times N}\sum_{i=1}^{N^2} w(t_i) \cdot x_i, \tag{5}$$

where i is the position index in the $N \times N$ grid to be pooled corresponding to position p. t_i and x_i are the values of the weighted map T and input feature map x at position i. w means the softmax normalization operation.

The feature map generated from Wei-Pooling is fed into SRM for long-range dependency modeling (upper branch) and local independency modeling (bottom branch). In the upper branch, the input feature C' is processed by a 3×3 convolution layer and three independent 1×1 convolution layers to generate three new feature map Q, K and V, where $\{Q, K, V\} \in \mathbb{R}^{C\times \frac{H}{N}\times \frac{W}{N}}$. Then they are all reshaped to $\mathbb{R}^{C\times N'}$, where $N' = \frac{H}{N} \times \frac{W}{N}$. After that, a similarity matrix S is calculated to indicate the correlation between any two pixels in Q and K. To compensate for the deficiency of global semantic information, S is used as the weight indicator to aggregate the global information and then element-wise added to the input feature C' to get the enhanced feature map. This process can be formulated as follows:

$$S = softmax((W_Q\boldsymbol{x})^T \times W_K\boldsymbol{x}), \tag{6}$$

$$\boldsymbol{y} = \boldsymbol{x} + S \times (W_V\boldsymbol{x}), \tag{7}$$

where W_Q, W_K, and W_V refer to three independent embedding matrices. "$\times$" denotes the matrix multiplication. Compared with general convolution layers that only focus on local information, this mechanism aggregates the distinguishable information of all pixels to highlight the feature of text region. Moreover, inspired by R-Net [37], we further use a parallel branch to highlight the local distinguish information. As shown in the lower branch in SRM, an attention map P is generated by a 3×3 convolution layer. Then we element-wise multiply P and the original feature map C' to highlight features. The enhanced features from two branches are element-wise added and processed by a de-convolution layer with stride N to restore the feature map back to the same resolution as C_5. Finally, we perform a residual operation to fuse the enhanced feature and original feature map C_5 to obtain the final feature

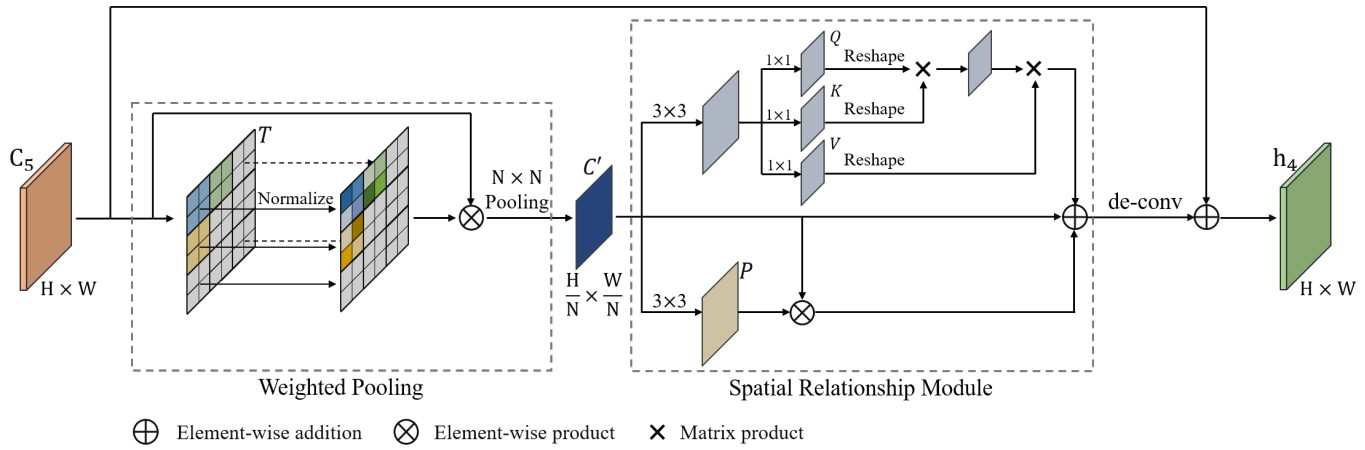


Fig. 4. The architecture of Efficient Spatial Relationship Module. It is a residual network with Weighted Pooling and Spatial Relationship Module. The upper and lower branches of Spatial Relationship Module are used to obtain global contextual information and local structural information respectively.

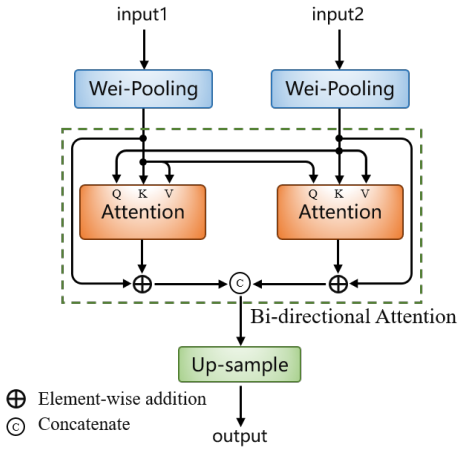


Fig. 5. Relationship Linking Module using Bi-directional Attention.

map h_4. Specially, we set N to 2 in both Wei-Pooling and SRM in this paper.

Different from the conventional attention mechanism [9], [10], [11], [12], ESRM mainly focuses on reducing the amount of computation without losing much information. There are two similar ways to improve efficiency: max pooling and average pooling. Max pooling only retains the max value in a block and causes information loss in the context. Average pooling does not take the importance of features at each location into account. To preserve the information in the context and maintain the highlight of the distinguishing features, the proposed Wei-Pooling integrates the sampled pixels with learned weights, which is quite different from existing pooling approaches and cannot be replaced. Benefiting from the Wei-Pooling, ESRM effectively captures the context information while maintaining high efficiency.

D. Adaptive Dilation Module

As shown in Fig. 2, ADM takes the fused feature map F as input and outputs three maps: text region map, kernel map, and dilation map. Text region map and kernel map are probability maps, where the value indicates the probability of this pixel inside the text region or kernel region. However, as the kernel map only focuses on the center region of text, the kernel branch working alone cannot fully perceive the text feature, suffering from the incomplete detection results and false negatives. Thus, we adopt the scheme of parallel work of these two branches. Specifically, the feature F is fed into both the top and bottom branches to predict the text region map and kernel map, which contain a 3×3 convolution layer and two de-convolution layers. Because these two branches are independent of each other, they can generate the expected segmentation map based on the feature F under different supervision. Text region prediction branch working supplementarily to the kernel map generation has also proved beneficial in the ablation experiment.

Besides, as shown in Eq. (2), considering that the variable dilation factor is related to both kernel region and text region, we take the feature from two branches into account and use a Relationship Linking Module (RLM) to calculate the inner relationship from two branches for each text instance. Then the dilation map is obtained by processing the fused feature from RLM with two de-convolution layers to restore the size back to the original resolution. In the inference stage, the original text region can be easily reconstructed from the kernel region using the unique dilation factor in the dilation map.

Concretely, two ways to implement RLM are considered for this process:

- **Concatenate + convolution layer.** Simply concatenate the feature from two branches and use a $k \times k$ convolution layer to fuse two feature maps. Different k is experimented in the ablation study.
- **Bi-directional Attention.** As shown in Fig. 5, to reduce the extra computation cost, the two feature maps are first down-sampled through Wei-Pooling (described in Sec.III.C). Then, two feature maps are used as Query vectors in attention mechanism to calculate the relationship with each other, respectively. The Attention module in Fig. 5 has the same structure as the upper branch in SRM. The output of Bi-directional Attention is the concatenated

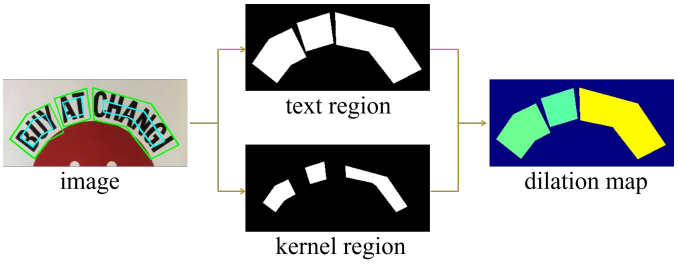


Fig. 6. Label generation. Different colors in the dilation map indicate different dilation factors.

feature of two branches after residual connection. Finally, up-sample the result from Bi-directional Attention to obtain the fused feature.

With introducing little computation cost, the RLM can be easily embedded into the network. In our experiments, they all achieve promising results.

E. Training Objective

1) Aspect ratio-aware dilation factor generation: As depicted in Fig. 6, the kernel region is generated by shrinking the original polygon G, formed by a set of points, to shrunk polygon G_k with distance D using the Vatti clipping algorithm [34].

$$D = \frac{R_s \cdot S_g}{C_g} = \frac{(1-r^2)\cdot S_g}{C_g}, \tag{8}$$

where S_g and C_g are the area and perimeter of G. r is the shrink factor. In our implementation, we set $r = 0.4$.

The procedure of dilation map label generation is an inverse of the kernel map generation process. Pixels inside an instance region are all set to R, which is calculated as

$$R = \frac{D \cdot C_k}{S_k}, \tag{9}$$

where D is the distance in Eq. (8), C_k and S_k are the perimeter and area of G_k. As the dilation map in Fig. 6 shows, each instance with a different aspect ratio has its own unique dilation factor. The generated label is used to guide the prediction of the dilation map.

2) Optimization: The loss function of the network is a weighted sum of each branch in the prediction stage, which can be expressed as

$$L = L_g + \alpha L_k + \beta L_d, \tag{10}$$

where L_g, L_k and L_d denote text region map loss, kernel map loss and dilation map loss. α and β are hyper-parameters to balance the numeric values of each branch. In our experiment, α and β are both set to 5.

We adopt a binary cross-entropy loss for L_g. On account of the imbalance between positive pixels and negative pixels, hard negative mining is applied to avoid the bias with the ratio of positives and negatives setting 1:3.

$$L_g = -\frac{1}{N}[\sum_{i\in S} y_i log\hat{y}_i + (1-y_i)log(1-\hat{y}_i)], \tag{11}$$

where S and N are the sampled set and the number of pixels in S.

L_k is optimized with a dice loss to achieve a robust performance on scale variance.

$$L_k = 1 - \frac{2\,|X \cap Y|}{|X| + |Y|}, \tag{12}$$

where $X \cap Y$ is the intersection between X and Y, and $|X|$ means the sum of the pixels in X.

L_d is computed as the sum of L1 distances between the predicted dilation map and label in the text region. It can be expressed as

$$L_d = \sum_{i\in S_k} |\hat{y}_i - y_i|, \tag{13}$$

where S_k is the set of text regions.

The dice loss is a scale-invariant loss function based on the IoU of two regions. It has good performance for small size region (shrunk region). However, dice loss lays emphasis on the overlapping area of two regions instead of the detail of each pixel. The BCE loss can capture pixel-level differences. To obtain more detailed information of the text region for enhancing the sensitivity of text region, we adopt BCE loss in the text region branch. According to the merits of these two loss functions, the parallel structure with different loss functions makes the backbone learn the shared feature better while still considering the characteristic of target maps.

F. Inference

In the inference period, only the kernel prediction map and dilation prediction map are needed to reconstruct the text instance. The bounding box can be obtained by the following steps: (1) Binarize the kernel prediction map with a fixed threshold. (2) Filter out kernels with low scores. (3) Average the coefficients of the predicted dilation map in each kernel to obtain an adaptive dilation factor R'. (4) Calculate the distance D' and use the Vatti clipping algorithm [34] to expand the kernel region. The calculation process of D' is formulated as following:

$$D' = \frac{R' \cdot S_k}{C_k}, \tag{14}$$

where S_k and C_k refer to the area and perimeter of the text kernel.

IV. Experiment

A. Datasets

The datasets used for experiments are introduced in the following:

1) SynthText: SynthText [38] is a synthetic dataset with about 800k images. These images are generated from 8k background natural scene images and 8 million rendered text instances. We only use this dataset to pre-train our model.

2) Total-Text: Total-Text [39] is a challenging dataset with arbitrary shape texts. It includes 1255 training images and 300 testing images. All text instances are labeled with several corner points at word level.

TABLE I
PERFORMANCE GAIN OF DILATION MAP. W/O MEANS WITHOUT DILATION MAP AND DETECTION RESULTS ARE OBTAINED BY SETTING R_d FIXED.

Dataset	Method	P	R	F	FPS
Total-Text	R_d = 2.0(w/o)	87.8	82.8	85.2	**18.0**
	R_d = 2.2(w/o)	87.4	82.5	84.9	
	R_d = 2.5(w/o)	86.4	81.5	83.9	
	ADNet	**90.6**	**84.4**	**87.4**	14.9
MSRA-TD500	R_d = 2.0(w/o)	86.1	71.3	78.0	**19.6**
	R_d = 2.2(w/o)	86.7	71.8	78.6	
	R_d = 2.5(w/o)	88.1	72.3	79.4	
	ADNet	**92.0**	**83.2**	**87.4**	16.7

3) MSRA-TD500: MSRA-TD500 [40] consists of 300 training images and 200 testing images with long text instances in multi-language. Following the previous works [1], [22], [29], [41], 400 more images from HUST-TR400 [42] are used for training.

4) CTW1500: CTW1500 [43] focuses on the curved text, including 1000 images for training and 500 images for testing. Text instances are all annotated with 14 vertices at sentence level.

5) ICDAR 2015: ICDAR 2015 [44] is proposed in Challenge 4 of ICDAR 2015 Robust Reading Competition. There are 1000 training images and 500 testing images. It provides word-level annotations with quadrilaterals.

B. Implementation Details

We use ResNet50 as our backbone. The training procedure can be divided into 2 steps: (1) Pre-train the network on SynthText dataset for 1 epoch. (2) Fine-tune the models on corresponding datasets for 1200 epochs. ADNet is trained on 1 GTX 1080 Ti GPU with a batch size 8 using SGD optimizer. The momentum is set to 0.9 and weight decay is set to 0.0001. For data augmentation, we implement random rotation with an angle from -10 to 10, random resizing from 0.5 to 2.0 and random flipping. Then all the training images are randomly cropped out a 640×640 patch from the above-processed images. The learning rate is initialized to 0.007 and adjusted in an exponential decay manner, that is, the current epoch learning rate is the initial learning rate multiplied by $(1 - \frac{cur\_epoch}{max\_epoch})^p$ where p is set to 0.9. Blurred texts labeled with DO NOT CARE are ignored in training period.

In this paper, baseline model is ADNet without ESRM, dilation map and text region map in ADM. In the inference period, we rescale the short side to a preset value (800 for Total-Text and CTW1500, 736 for MSRA-TD500 and 1136 for ICDAR 2015) and keep the aspect ratio unchanged. The original text region prediction is removed for better efficiency.

C. Ablation Study

We conduct several ablation studies on Total-Text [39] and MSRA-TD500 [40] to prove the effectiveness of ADNet.

TABLE II
PERFORMANCE GAIN OF SUPERVISION ON THE ORIGINAL TEXT REGION. W/O MEANS WITHOUT.

Dataset	Method	P	R	F
Total-Text	w/o text region	89.1	81.7	85.2
	ADNet	**90.6**	**84.4**	**87.4 (+2.2)**
MSRA-TD500	w/o text region	88.4	82.5	85.3
	ADNet	**92.0**	**83.2**	**87.4 (+2.1)**

1) Discussion about Adaptive Dilation Module: First, we evaluate the benefits of dilation map on Total-Text [39] and MSRA-TD500 [40]. As shown in Tab. I, for the reconstruction with constant dilation factor, the value of R_d degenerates into a hyper-parameter selection problem, which lacks theoretical support. At the cost of about 3 FPS, our ADNet equipped with dilation map can boost the F-measure by 2.2% for Total-Text and 8.0% for MSRA-TD500.

As Eq. (3) and Eq. (4) show, the dilation factor is related to the aspect ratio of a polygon. Text instances in both datasets have a wide range of aspect ratios. With an adaptive dilation factor for texts in different aspect ratios, our ADNet achieves higher precision, recall and F-measure on both datasets. Moreover, because MSRA-TD500 [40] contains more long text instances, the learned dilation map brings a much higher increase. The visualization results are shown in the following.

Then, we discuss the complementary work of text region branch. It is worth noting that the kernel region is crucial for separating adjacent texts, which cannot be abandoned. But the supervision of kernel region only focuses on a part of the text region, which is not conducive for the model to detect the whole text region, leading to false negatives and low precision of the detection results. Therefore, we add a new branch to learn the original text region. The two branches need a shared feature sensitive to text instance and a specific feature to obtain the original/shrunk region. The shared feature, like the texture feature of text instance, is extracted by the backbone and both branches can play a role in the optimization of backbone. The supplementary supervision on text region map increases the model's perception of the full-text region, not just the shrunk part of a text instance, which is beneficial for the model to detect the text more completely. In Tab. II, we can see that with the supervision on the original text region, the F-measure increases 2.2% for Total-Text [39] and 2.1% for MSRA-TD500 [40]. Precision and recall are both improved, which indicates that the added text region branch effectively reduces error samples and false negatives.

2) Choice of Relationship Linking Module: As Tab. III shows, the function of employing concatenation with 3×3 convolution and without Wei-Pooling achieves a slightly better result than others. As shown in the first three lines, when simply using a convolution layer to process the cascaded features, convolution kernels of different sizes have different performance. The 1×1 filter has limited receptive field and does not fully consider the correlation information contained in the surrounding position. Only focusing on current pixel information also introduces noises, which is unfavorable to the final dilate map because the dilation factor is related to all the

TABLE III
COMPARISON OF DIFFERENT RELATIONSHIP LINKING FUNCTIONS ON TOTAL-TEXT. WEI-POOLING MEANS WHETHER EQUIPPED WITH WEI-POOLING AND UP-SAMPLE LAYER AT BOTH ENDS OF THE FUNCTION.

Function	Wei-Pooling	P	R	F	FPS
concat + 1×1 conv	$\times$	90.0	84.2	87.0	**14.9**
concat + 3×3 conv	$\times$	90.6	**84.4**	**87.4**	**14.9**
concat + 5×5 conv	$\times$	**90.7**	83.6	87.0	**14.9**
independent 3×3 conv	$\times$	90.4	83.0	86.5	**14.9**
concat + 3×3 conv	✓	90.4	84.2	87.2	14.7
Bi-directional Attention	✓	90.6	83.5	86.9	13.2

TABLE IV
ABLATION EXPERIMENTS OF ESRM ON TOTAL-TEXT. $N \times N$ CONV MEANS USING GENERAL CONVOLUTION LAYER TO DOWN-SAMPLE.

ESRM	Down-sampling	P	R	F	FPS
-	-	89.7	82.3	85.9	**15.3**
✓	-	**90.9**	84.3	**87.5**	13.5
✓	2×2 conv	90.0	83.3	86.5	14.9
✓	3×3 conv	89.4	84.2	86.7	14.9
✓	Average Pooling	90.1	83.9	86.9	15.2
✓	Max Pooling	90.6	82.4	86.3	15.2
✓	**Wei-Pooling**	90.6	**84.4**	87.4	14.9

pixels inside a text region. The 5×5 filter has a larger receptive field than 3×3, and the neighbor information considered is almost three times that of the latter. However, the information in 3×3 region is enough to model the correlation between two feature maps. Most of the information in 5×5 filter is redundant and hampers the learning of network.

Moreover, the second and last two lines indicate that a 3×3 filter is capable of modeling the correlation of two feature maps and obtaining incredible results. Although our devised Bi-directional Attention can theoretically reason the relationship in a more complicated way, the increased complexity makes the network difficult to optimize. This phenomenon conforms to the principle of Occam's razor, which believes that the simple network is more reliable than the complex one. However, when compared with independent 3×3 filter that directly predicts the dilation map from feature map F, Bi-directional Attention shows the superiority in considering the correlation of two feature maps. Therefore, we believe that the proposed Bi-directional Attention can provide insights for modeling correlations in other cases. Similarly, the added Wei-Pooling and up-sample layer at the both ends of the concatenation and convolution layer bring a slight decline in performance. We are inclined to think that a convolution layer is simple and effective enough to model the correlation here. Therefore, the added Wei-Pooling layer brings redundant parameters and computation, causing the slight drop in accuracy and speed.

3) Discussions about Efficient Spatial Relationship Module: ESRM consists of SRM and Wei-Pooling. As the first two lines in Tab. IV show, ESRM without down-sampling, namely SRM, brings 1.6% improvement in F-measure on Total-Text [39] by capturing the global contextual information and local structural information. We also evaluate the benefits and losses of different down-sampling methods. As shown in the other lines in Tab. IV, down-sampling with general convolution layer, average pooling or max pooling brings about a 1.4 FPS improvement in speed, but at the cost of at least 0.6% decrease in F-measure due to the information loss. To restore more information in pooling procedure, the proposed Wei-Pooling increases the speed by 1.4 FPS with only subtle loss of accuracy, indicating almost no information loss in Wei-Pooling. Overall, our ESRM only causes a 0.4 FPS decrease but achieves a remarkable improvement of 1.5% F-measure on Total-Text.

TABLE V
THE GAIN OF F-MEASURE OF ESRM ON TOTAL-TEXT IN DIFFERENT ASPECT RATIO RANGES.

	[1/3, 3]	[1/3, 1/5]&[3, 5]	else
w/o ESRM	88.2	86.4	74.0
ADNet	89.0(+0.8)	87.8(+1.4)	77.0(+3.0)

TABLE VI
PERFORMANCE ON TOTAL-TEXT WITH DIFFERENT LOSS FUNCTION. BCE AND DICE MEAN BINARY CROSS-ENTROPY LOSS AND DICE LOSS.

L_g	L_k	P	R	F
BCE	Dice	**90.6**	**84.4**	**87.4**
BCE	BCE	87.8	61.2	72.1
Dice	Dice	88.9	83.1	85.9
Dice	BCE	89.4	62.2	73.4

Furthermore, as shown in Tab. V, the gain of ESRM increases from 0.8% to 3.0% as the value of aspect ratio increases, which effectively indicates that ESRM is helpful in tackling text instances with extreme aspect ratios.

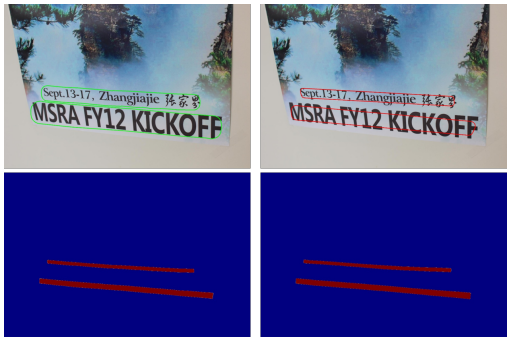
4) Choice of loss functions: As shown in Tab. VI, ADNet performs best when text region branch loss L_g and kernel branch loss L_k adopt BCE loss and dice loss, respectively. For kernel branch, when replacing dice loss with BCE loss, recall drops significantly because the predicted small text kernel regions have low confidence scores. Substituting BCE loss for dice loss in text region branch also leads to performance degradation. This is because the dice loss is a scale-invariant loss that focuses on the intersection of two regions rather than the detail of each pixel. The BCE loss tends to capture pixel-level differences. The adoption of dice loss in kernel branch can improve the confidence of kernel region and reduce the occurrence of false negative samples. At the same time, the small error in kernel region prediction will be compensated in the subsequent adaptive dilation process, so that it does not need to pay much attention to the boundary details. The text region branch aims to enhance the model's perception of text regions, so it is more suitable to adopt BCE loss.

5) Setting of shrink factor: Several experiments about the shrink factor r have been conducted in Tab. VII. It is a parameter used to trade off the network's ability to separate close texts and the ability to describe a text instance. Results show that when r is set to 0.4, ADNet has the best performance on balancing these two capabilities. Therefore, we use $r = 0.4$ in all the following experiments.

6) Qualitative analysis: Fig. 7, 8, 9 are the visualizations of ADNet with and without dilation map, text region, and

TABLE VII
PERFORMANCE ON TOTAL-TEXT WITH DIFFERENT SETTING OF SHRINK FACTOR.

Shrink factor	P	R	F
0.2	88.8	81.3	84.9
0.4	**90.6**	**84.4**	**87.4**
0.6	89.4	83.4	86.3
0.8	88.3	79.6	83.7



(a) ADNet (b) w/o dilation map

Fig. 7. Detection results with and without dilation map. Bottom pictures are the kernel maps. ADNet generates more accurate results with almost the same kernel map.

(a) ADNet (b) w/o text region

Fig. 8. Detection result with and without text region branch. Top samples show the improvement for precision and bottom for recall.

(a) ADNet (b) w/o ESRM

Fig. 9. Detection result with and without ESRM. w/o means without.

TABLE VIII
COMPARISON WITH STATE-OF-THE-ART METHODS ON TOTAL-TEXT. PRE-TRAINED ON DATASET OTHER THAN SYNTHTEXT IS NOT INCLUDED.

Method	Venue	P	R	F
TextSnake [29]	ECCV'18	82.7	74.5	78.4
TextField [4]	TIP'19	81.2	79.9	80.6
Dai *et al.* [45]	TMM'19	84.6	78.6	81.5
PSE-1s [2]	CVPR'19	84.0	78.0	80.9
LOMO [23]	CVPR'19	88.6	75.7	81.6
CRAFT [46]	CVPR'19	87.6	79.9	83.6
PAN [30]	ICCV'19	89.3	81.0	85.0
MSR [25]	IJCAI'19	83.8	74.8	79.0
SPCNET [47]	AAAI'19	83.0	82.9	82.9
DB [1]	AAAI'20	87.1	82.5	84.7
Boundary [48]	AAAI'20	85.2	83.5	84.3
OPMP [49]	TMM'20	87.6	82.7	85.1
TextRay [24]	MM'20	83.5	77.9	80.6
ABCNet [50]	CVPR'20	87.9	81.3	84.5
DRRG [51]	CVPR'20	86.5	**84.9**	85.7
ContourNet [52]	CVPR'20	86.9	83.9	85.4
Mask-TTD [53]	TIP'20	79.1	74.5	76.7
Dai *et al.* [54]	TMM'21	86.7	82.6	84.6
PCR [55]	CVPR'21	88.5	82.0	85.2
FCENet [56]	CVPR'21	89.3	82.5	85.8
TextBPN [57]	ICCV'21	90.3	84.7	**87.4**
DBNet++ [58]	TPAMI'22	88.9	83.2	86.0
ADNet	-	**90.6**	84.4	**87.4**

ESRM, respectively. The results in Fig. 7 show that the kernel regions are almost identical with and without the aspect ratio-aware dilation map prediction branch. However, the detection results are quite different after the dilation process, which demonstrates the effectiveness of our proposed adaptive dilation factor. As the samples in Fig. 8 show, the addition of the text region branch effectively depresses the false negatives and completely detects the text instances. And we can see from Fig. 9 that ADNet is prone to detect long and thin text instances into multiple small ones without ESRM. These visualization results further illustrate the effectiveness of our method.

D. Comparisons with Previous Methods

In this section, we compare ADNet with the existing state-of-the-art methods on four benchmark datasets, including two curved text datasets, a multi-language dataset, and a multi-oriented dataset. We employ IoU@0.5 as our evaluation protocol, which is the same as DB [1].

1) Evaluation on Curved Text Benchmark: We evaluate ADNet on Total-Text to test the performance of detecting curved texts. As Tab. VIII shows, our method achieves results of 90.6%, 84.4% and 87.4% in precision, recall and F-measure. Compared to the boundary representation methods [23], [25], [48], [24], [50], [55], ADNet has an F-measure of 2.2% higher than the best method among them. As for the whole text region segmentation-based methods, TextField [4], CRAFT [46] and SPCNET [47], ADNet surpasses them by a large margin (87.4% vs 80.6%, 83.6% and 82.9% in F-measure). We attribute this superior performance to the kernel map in Adaptive Dilation Module. Generally, in the segmentation task, the internal area of the target has a high confidence and is easier to be detected. Therefore, the whole text region segmentation map with low border region confidence used in [4], [46], [47] is not as accurate as the kernel map to represent a text instance.

TABLE IX
COMPARISON WITH STATE-OF-THE-ART METHODS ON CTW1500. PRE-TRAINED ON DATASET OTHER THAN SYNTHTEXT IS NOT INCLUDED.

Method	Venue	P	R	F
CTPN [18]	ECCV'16	60.4	53.8	56.9
EAST [22]	CVPR'17	78.7	49.1	60.4
TextSnake [29]	ECCV'18	67.9	**85.3**	75.6
CTD+TLOC [59]	PR'19	77.4	69.8	73.4
TextField [4]	TIP'19	83.0	79.8	81.4
Dai *et al.* [45]	TMM'19	85.7	85.1	85.4
PSE-1s [2]	CVPR'19	84.8	79.7	82.2
Tian *et al.* [28]	CVPR'19	82.7	77.8	80.1
Wang *et al.* [60]	CVPR'19	80.1	80.2	80.1
CRAFT [46]	CVPR'19	86.0	81.1	83.5
PAN [30]	ICCV'19	86.4	81.2	83.7
MSR [25]	IJCAI'19	85.0	78.3	81.5
TextRay [24]	MM'20	80.2	78.4	79.3
R-Net [37]	TMM'20	74.6	71.0	72.8
OPMP [49]	TMM'20	85.1	80.8	82.9
DB [1]	AAAI'20	86.9	80.2	83.4
ABCNet [50]	CVPR'20	84.4	78.5	81.4
DRRG [51]	CVPR'20	85.9	83.0	84.5
ContourNet [52]	CVPR'20	83.7	84.1	83.9
Mask-TTD [53]	TIP'20	79.7	79.0	79.4
Dai *et al.* [54]	TMM'21	87.2	81.7	84.4
PCR [55]	CVPR'21	87.2	82.3	84.7
FCENet [56]	CVPR'21	87.6	83.4	85.5
TextBPN [57]	ICCV'21	87.8	81.5	84.5
DBNet++ [58]	TPAMI'22	87.9	82.8	85.3
ADNet	-	**88.2**	83.1	**85.6**

TABLE X
COMPARISON WITH STATE-OF-THE-ART METHODS ON MSRA-TD500. PRE-TRAINED ON DATASET OTHER THAN SYNTHTEXT IS NOT INCLUDED.

Method	Venue	P	R	F
EAST [22]	CVPR'17	87.3	67.4	76.1
Corner [41]	CVPR'18	87.6	76.2	81.5
TextSnake [29]	ECCV'18	83.2	73.9	78.3
PixelLink [3]	AAAI'18	83.0	73.2	77.8
TextField [4]	TIP'19	87.4	75.9	81.3
CRAFT [46]	CVPR'19	88.2	78.2	82.9
PAN [30]	ICCV'19	84.4	**83.8**	84.1
MSR [25]	IJCAI'19	87.4	76.7	81.7
Tian *et al.* [28]	CVPR'19	84.2	81.7	82.9
Mask TextSpotter-v2 [61]	TPAMI'19	80.8	68.6	74.2
Mask TextSpotter-v3 [33]	ECCV'20	90.7	77.5	83.5
R-Net [37]	TMM'20	83.7	79.7	81.7
OPMP [49]	TMM'20	86.0	83.4	84.7
DB [1]	AAAI'20	91.5	79.2	84.9
DRRG [51]	CVPR'20	88.0	82.3	85.1
PCR [55]	CVPR'21	90.8	83.5	87.0
TextBPN [57]	ICCV'21	85.4	80.7	83.0
DBNet++ [58]	TPAMI'22	91.5	83.3	87.2
ADNet	-	**92.0**	83.2	**87.4**

Particularly, as for the shrunk polygon-based methods [29], [2], [31], [30], [1], [58], ADNet outperforms them by at least 1.4% in F-measure. The improvement mainly ascribes to the dilation map in ADM and Efficient Spatial Relationship Module. The former considers the relationship between text region and kernel region, which is the basis of generating tighter bounding boxes. The latter helps ADNet capture global contextual information to handle texts with extreme aspect ratios. It is worth mentioning that the most significant difference between ADNet and DB [1] is the introduction of aspect ratio-aware dilation map. The adaptive dilation factor brings a 2.7% gap in F-measure, which confirms the effectiveness of our method. In addition, ADNet outperforms DBNet++ [58] by 1.4% in F-measure. This is because the Adaptive Scale Fusion (ASF) in DBNet++ [58] is designed to compute channel attention and generate robust representations of multi-scale text instances. Compared to our ESRM, which focuses on spatial attention, ASF still lacks the ability to capture long-range dependencies, leading to worse detection results for text instances with extreme aspect ratios.

2) Evaluation on Long Curved Text Benchmark: To show the performance for detecting long curved texts, we evaluate ADNet on CTW1500 [43]. As shown in Tab. IX, ADNet achieves a new state-of-the-art performance on CTW1500 [43] with F-measure of 85.6%. Compared with the regression-based methods [22], [24], [55], ADNet surpasses them by at least 0.9% due to the accurate text region prediction map. Though the top-down method PCR [55] iteratively refines the boundary points to detect long texts and generates tighter boundaries, the detection results are lower than that of ADNet (84.7% vs 85.6%). Also, compared with Mask-TTD [53], which is designed for tighter bounding boxes, ADNet outperforms it by a large margin (79.4% vs 85.6%). This notable gap demonstrates the ability of ESRM to perceive long text instances and ADM to obtain more compact bounding boxes.

Similarly, the shrunk polygon-based methods [1], [2], [29], [30] can only reach a maximum F-measure of 83.7%. ADNet achieves a better result with 1.9% improvement in F-measure benefiting from the adaptive dilation factor and ESRM. The texts in CTW1500 [43] are labeled at the sentence level and are more likely to be long texts. Therefore, ESRM, aiming to capture global contextual information, is very helpful. Specifically, PSENet [2] uses a series of kernels to describe a text instance and progressively merges them into one. Learning too many branches makes it hard to balance the optimization process of each, which results in an inaccurate text region. Our ADNet using only two kernels (shrunk region and text region) achieves a much higher performance (3.4% in F-measure) than PSENet [2]. Besides, ADNet outperforms DB [1], which uses a fixed dilation factor, by 2.2% in F-measure, which confirms the effectiveness of the adaptive dilation factor.

3) Evaluation on Long Oriented Text Benchmark: We also prove the robustness of our method on the multi-language and multi-oriented dataset MSRA-TD500. As shown in Tab. X, ADNet achieves a result of 92.0%, 83.2% and 87.4% in precision, recall and F-measure, which is superior to the previous state-of-the-art method PCR [55] with 0.4% increase in F-measure. Compared with methods designed for multi-oriented texts [62], [22], [41], our method achieves the improvement of at least 4.4%, 7.0%, and 5.9% in precision, recall, and F-measure, respectively. Some similar shrunk polygon-based methods, PAN [30], MSR [25], and DB [1], reach the 84.1%, 81.7%, and 84.9% in F-measure, which are all lower than ADNet. In particular, DB [1] adopts a fixed dilation factor, and there is no global context information capture mechanism like ESRM in ADNet. Using the same backbone as DB [1], ADNet promotes the F-measure from 84.9% to 87.4%.

4) Evaluation on Oriented Text Benchmark: We evaluate the capacity of our method on ICDAR 2015 for detecting multi-oriented texts. Evaluation results are listed in Tab. XI.

Fig. 10. Some experimental results on text instances in various shapes, including curved text, multi-oriented text, long text instance and multi-language text. For each unit, the top right is the kernel map, and the bottom right is the dilation map.

TABLE XI
COMPARISON WITH STATE-OF-THE-ART METHODS ON ICDAR2015. MULTI-SCALE TESTING AND ENSEMBLE ARE NOT INCLUDED.

Method	Venue	P	R	F
CTPN [18]	ECCV'16	74.2	51.6	60.9
EAST [22]	CVPR'17	83.6	73.5	78.2
TextSnake [29]	ECCV'18	84.9	80.4	82.6
PixelLink [3]	AAAI'18	85.5	82.0	83.7
TextField [4]	TIP'19	84.3	83.9	84.1
Dai *et al.* [45]	TMM'19	86.2	82.7	84.4
PSE-1s [2]	CVPR'19	86.9	84.5	85.7
LOMO [23]	CVPR'19	91.3	83.5	87.2
CRAFT [46]	CVPR'19	89.8	84.3	86.9
Tian *et al.* [28]	CVPR'19	88.3	85.0	86.6
Wang *et al.* [60]	CVPR'19	90.4	83.3	86.8
PAN [30]	ICCV'19	82.9	77.8	80.3
MSR [25]	IJCAI'19	86.6	78.4	82.3
SPCNET [47]	AAAI'19	88.7	85.8	87.2
R-Net [37]	TMM'20	88.7	82.8	85.6
OPMP [49]	TMM'20	89.1	85.5	87.3
DB [1]	AAAI'20	91.8	83.2	87.3
Boundary [48]	AAAI'20	82.2	**88.1**	85.0
DRRG [51]	CVPR'20	88.5	84.7	86.6
ContourNet [52]	CVPR'20	87.6	86.1	86.9
Dai *et al.* [54]	TMM'21	87.2	81.3	84.1
TextMountain [27]	PR'21	89.5	83.1	86.2
FCENet [56]	CVPR'21	90.1	82.6	86.2
DBNet++ [58]	TPAMI'22	90.9	83.9	87.3
ADNet	-	**92.5**	83.7	**87.9**

TABLE XII
EXPERIMENTS ON CTW1500, TOTAL-TEXT AND MSRA-TD500 WITH HIGH IOU THRESHOLDS. * INDICATES RESULTS FROM [53].

Dataset	IoU	Method	P	R	F
CTW1500	0.6	CTD+TLOC* [59]	67.7	61.3	64.3
		Mask TTD* [53]	68.4	74.5	71.3
		ADNet	**85.4**	**80.5**	**82.9**
	0.7	CTD+TLOC* [59]	49.1	44.4	46.6
		Mask TTD* [53]	57.1	62.2	59.5
		ADNet	**78.7**	**74.2**	**76.4**
	0.8	CTD+TLOC* [59]	20.6	18.6	19.5
		Mask TTD* [53]	33.9	36.9	35.3
		ADNet	**56.1**	**53.1**	**54.6**
Total-Text	0.6	DB [1]	81.5	75.7	78.5
		ADNet	**85.1**	**80.5**	**82.7**
	0.7	DB [1]	69.1	64.5	66.7
		ADNet	**76.1**	**72.2**	**74.1**
	0.8	DB [1]	37.2	35.0	36.1
		ADNet	**52.6**	**50.2**	**51.4**
MSRA-TD500	0.6	DB [1]	87.4	77.5	82.1
		ADNet	**88.0**	**80.8**	**84.2**
	0.7	DB [1]	74.0	65.6	69.6
		ADNet	**82.6**	**75.8**	**79.0**
	0.8	DB [1]	41.1	36.4	38.6
		ADNet	**68.6**	**62.9**	**65.6**

TABLE XIII
INFERENCE SPEED AND F-MEASURE. "-" INDICATES NO OFFICIAL MODEL PROVIDED.

Method	Total-Text		CTW1500		MSRA-TD500		ICDAR2015	
	F	FPS	F	FPS	F	FPS	F	FPS
PSE-4s [2]	-	-	79.9	10.0	-	-	84.9	4.2
PSE-1s [2]	-	-	82.2	3.9	-	-	85.7	1.4
DB [1]	84.7	**18.0**	-	-	84.9	**19.1**	87.3	**7.0**
ContourNet [52]	85.4	3.8	83.9	4.5	-	-	86.9	3.5
ADNet	**87.4**	14.9	**85.6**	**14.2**	**87.4**	16.7	**87.9**	5.7

ADNet reaches an excellent performance at 92.5%, 83.7% and 87.9% for precision, recall and F-measure, outperforming previous state-of-the-art method DB [1] by 0.6% in F-measure. Compared with the shrunk polygon-based methods [1], [2], [27], [30], [25], ADNet achieves a higher precision by at least 0.7%, indicating the more complete and tight detection results. We attribute this improvement to the adaptive guidance of aspect ratio-wise dilation factor for each text instance, which takes the correlation between original and shrunk region into account.

Accordingly, from the experiment results on these benchmarks, ADNet can achieve state-of-the-art performance on several datasets with a promising speed. Detection results of text instances with various shaped are visualized in Fig. 10.

E. Discussion about the tightness of bounding boxes

To further demonstrate the effectiveness of our method to generate more accurate and tight boundaries, we evaluate ADNet under stricter evaluation metrics. Besides, we reimplement DB [1] and use the official code and model to evaluate performance on Total-Text [39] and MSRA-TD500 [40]. As indicated in Tab. XII, the precision, recall, and F-measure of ADNet are all remarkably higher than the previous methods [1], [53], [59] on all three datasets. Notably, the more stringent the evaluation metric is, the greater the gap between ADNet and other methods [1], [53], [59] is, which strongly demonstrates the extraordinary performance of our method in generating more compact boundaries.

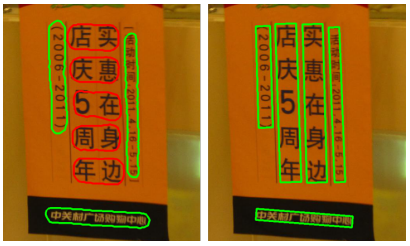


Fig. 11. Failure case. Left is the result from ADNet, and right is the ground-truth.

F. Discussion about the speed

The inference speed is correlated to many things, not only the GPU but also CPU and the cost of parallel tasks. When the models are tested in the same environment, does the comparison make sense. We reimplement the official code and model to test on 1 NVIDIA TITAN X GPU in the same environment for fairness. As depicted in Tab. XIII, ADNet can reach the top performance on CTW1500 [43] in the F-measure of 85.6% and speed at 14.2 FPS. On the other three benchmarks, compared with the most popular method DB [1], the proposed ADNet can obtain a much higher F-measure with a small speed loss, achieving a better trade-off between accuracy and efficiency. We attribute the increase of F-measure and the speed loss to the proposed adaptive dilation factor and ESRM. By capturing long-range dependencies and using instance-wise dilation factor instead of a fixed one, ADNet can generate a more complete and tight detection results.

G. Limitation

Our method can accurately detect a single column of vertical text but fails to detect two text lines that exist side by side vertically. As shown in Fig. 11, ADNet tends to describe them as a series of horizontal texts instead of two vertical texts. The same problem also exists in many segmentation-based methods [1], [2], [25]. Due to the lack of these training samples in the dataset, the model is not robust to this type of text instances. ADNet obtains detection results based only on visual information, without considering semantic information. This limitation can be improved in the end-to-end text spotting task, which takes the semantic information into account.

V. Conclusion

In this paper, we propose a novel framework named ADNet for arbitrary shape scene text detection. The Adaptive Dilation Module considers the relationship between the shrunk and original region of a text instance to obtain a aspect ratio-aware dilation factor. Besides, an Efficient Spatial Relationship Module is introduced to obtain long-range dependencies with little computation. Benefiting from these two modules, ADNet can generate a more compact and complete text region in various aspect ratios at an impressive speed. Experiments on several benchmarks demonstrate that our method achieves state-of-the-art performance. As for future work, we will combine our detection framework with a recognition branch for an end-to-end text spotting system and further improve the performance.

Appendix
Derivation of dilation factor

Kernel region is obtained by shrinking the polygon G with distance D using the Vatti clipping algorithm [34]. D is calculated as

$$D = \frac{R_s \cdot S_g}{C_g}, \tag{15}$$

where S_g and C_g are the area and perimeter of G. R_s is the shrink factor.

Similar to the process of shrinking, when reconstructing original region from shrunk region, the dilation distance is calculated as

$$D = \frac{R_d \cdot S_k}{C_k}. \tag{16}$$

S_k and C_k are the area and perimeter of shrunk polygon. R_d is the dilation factor.

Dilation and shrinking distance should be equal, namely

$$\frac{R_s \cdot S_g}{C_g} = \frac{R_d \cdot S_k}{C_k}. \tag{17}$$

Without losing generality, we assume the original polygon G as an ellipse with the major axis of a and minor axis of b. Then, the major and minor axes of the shrunk region are $a - D$ and $b - D$. Following the ellipse's area calculation formula $S = \pi ab$ and perimeter formula $C = 2\pi b + 4(a - b)$, Eq. (17) can be expressed as

$$R_s \frac{\pi ab}{2\pi b + 4(a-b)} = R_d \frac{\pi(a-D)(b-D)}{2\pi(b-D) + 4(a-b)}, \tag{18}$$

$$R_d = R_s \frac{a}{a-D} \frac{2 - \frac{4}{\pi} + \frac{4}{\pi}\frac{a-D}{b-D}}{2 - \frac{4}{\pi} + \frac{4}{\pi}\frac{a}{D}}. \tag{19}$$

For simplicity, let $c_1 = 2 - \frac{4}{\pi}$, $c_2 = \frac{4}{\pi}$ and $m = \frac{a}{b}$. Substituting Eq. (15) into the Eq. (19) gives:

$$R_d = R_s \frac{1}{c_1 + c_2 m - R_s}(c_1 + c_2 \frac{a-D}{b-D}) \tag{20}$$

$$= R_s \frac{1}{c_1 + c_2 m - R_s}[c_1 + c_2 \frac{m(c_1 + c_2 m - R_s)}{c_1 + c_2 m - m R_s}] \tag{21}$$

$$= R_s \frac{(c_1 + c_2 m)(c_1 + c_2 m - R_s) + c_1 R_s (1-m)}{(c_1 + c_2 m - R_s)(c_1 + c_2 m - m R_s)}. \tag{22}$$

Let $p = c_1 + c_2 m$, Eq. (20) reduces to

$$R_d = \frac{p R_s (p - R_s) + (2 - \frac{4}{\pi})(1-m) R_s}{(p - R_s)(p - m R_s)}. \tag{23}$$

When the polygon G is modeled as rectangle with width and height equal to a and b, Eq. (17) can be expressed as

$$R_s \frac{ab}{2(a+b)} = R_d \frac{(a-D)(b-D)}{2(a+b-2D)}. \tag{24}$$

With setting $m = \frac{a}{b}$, Eq. (24) reduces to

$$R_d = R_s \frac{m + 1 - 2\frac{D}{b}}{m+1} \frac{m}{(m - \frac{D}{b})(1 - \frac{D}{b})}. \tag{25}$$

Substituting Eq. (15) into the Eq. (25) gives:

$$R_d = R_s \frac{4(m+1)^2 - 4mR_s}{4(m+1)^2 - 2(m+1)^2 R_s + mR_s^2} \quad (26)$$

$$= \frac{-4R_s + 16}{\frac{m}{(m+1)^2}R_s^2 - 2R_s + 4} - 4. \quad (27)$$

Eq. (23) and Eq. (26) show that the dilation factor R_d is not only related to shrink factor R_s, but also to the aspect ratio of a text instance.

REFERENCES

[1] 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.

[2] 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 Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9336–9345.

[3] D. Deng, H. Liu, X. Li, and D. Cai, "Pixellink: Detecting scene text via instance segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[4] Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang, and X. Bai, "Textfield: learning a deep direction field for irregular scene text detection," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5566–5579, 2019.

[5] Y. Wang, H. Xie, S. Fang, J. Wang, S. Zhu, and Y. Zhang, "From two to one: A new scene text recognizer with visual language modeling network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 194–14 203.

[6] Y. Wang, H. Xie, S. Fang, M. Xing, J. Wang, S. Zhu, and Y. Zhang, "Petr: Rethinking the capability of transformer-based language model in scene text recognition," *IEEE Transactions on Image Processing*, vol. 31, pp. 5585–5598, 2022.

[7] F. Sheng, Z. Chen, and B. Xu, "Nrtr: A no-recurrence sequence-to-sequence model for scene text recognition," in *International Conference on Document Analysis and Recognition (ICDAR)*, 2019, pp. 781–786.

[8] Y. Du, Z. Chen, C. Jia, X. Yin, T. Zheng, C. Li, Y. Du, and Y.-G. Jiang, "Svtr: Scene text recognition with a single visual model," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 2022, pp. 884–890.

[9] 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.

[10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[11] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[12] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 603–612.

[13] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 7, pp. 1480–1500, 2014.

[14] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 161–184, 2021.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[18] 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.

[19] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[20] 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.

[21] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 745–753.

[22] 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*, 2017, pp. 5551–5560.

[23] 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 Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 552–10 561.

[24] F. Wang, Y. Chen, F. Wu, and X. Li, "Textray: Contour-based geometric modeling for arbitrary-shaped scene text detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 111–119.

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

[26] 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*, 2015, pp. 3431–3440.

[27] Y. Zhu and J. Du, "Textmountain: Accurate scene text detection via instance segmentation," *Pattern Recognition*, vol. 110, p. 107336, 2021.

[28] Z. Tian, M. Shu, P. Lyu, R. Li, C. Zhou, X. Shen, and J. Jia, "Learning shape-aware embedding for scene text detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4234–4243.

[29] 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)*, 2018, pp. 20–36.

[30] 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 International Conference on Computer Vision*, 2019, pp. 8440–8449.

[31] W. Feng, W. He, F. Yin, X.-Y. Zhang, and C.-L. Liu, "Textdragon: An end-to-end framework for arbitrary shaped text spotting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9076–9085.

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

[33] M. Liao, G. Pang, J. Huang, T. Hassner, and X. Bai, "Mask textspotter v3: Segmentation proposal network for robust scene text spotting," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 706–722.

[34] B. R. Vatti, "A generic solution to polygon clipping," *Communications of the ACM*, vol. 35, no. 7, pp. 56–63, 1992.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[36] T.-Y. Lin, P. Dollár, 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*, 2017, pp. 2117–2125.

[37] Y. Wang, H. Xie, Z.-J. Zha, Y. Tian, Z. Fu, and Y. Zhang, "R-net: A relationship network for efficient and accurate scene text detection," *IEEE Transactions on Multimedia*, 2020.

[38] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2315–2324.

[39] C. K. Ch'ng and C. S. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 935–942.

[40] 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.

[41] 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, 2018, pp. 7553–7563.

[42] 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.

[43] L. Yuliang, J. Lianwen, Z. Shuaitao, and Z. Sheng, "Detecting curve text in the wild: New dataset and new solution," *arXiv preprint arXiv:1712.02170*, 2017.

[44] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu *et al.*, "Icdar 2015 competition on robust reading," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 1156–1160.

[45] P. Dai, H. Zhang, and X. Cao, "Deep multi-scale context aware feature aggregation for curved scene text detection," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 1969–1984, 2019.

[46] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9365–9374.

[47] 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, no. 01, 2019, pp. 9038–9045.

[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," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 160–12 167.

[49] S. Zhang, Y. Liu, L. Jin, Z. Wei, and C. Shen, "Opmp: An omnidirectional pyramid mask proposal network for arbitrary-shape scene text detection," *IEEE Transactions on Multimedia*, vol. 23, pp. 454–467, 2020.

[50] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9809–9818.

[51] 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*, 2020, pp. 9699–9708.

[52] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 753–11 762.

[53] Y. Liu, L. Jin, and C. Fang, "Arbitrarily shaped scene text detection with a mask tightness text detector," *IEEE Transactions on Image Processing*, vol. 29, pp. 2918–2930, 2019.

[54] P. Dai, Y. Li, H. Zhang, J. Li, and X. Cao, "Accurate scene text detection via scale-aware data augmentation and shape similarity constraint," *IEEE Transactions on Multimedia*, vol. 24, pp. 1883–1895, 2021.

[55] P. Dai, S. Zhang, H. Zhang, and X. Cao, "Progressive contour regression for arbitrary-shape scene text detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7393–7402.

[56] Y. Zhu, J. Chen, L. Liang, Z. Kuang, L. Jin, and W. Zhang, "Fourier contour embedding for arbitrary-shaped text detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3123–3131.

[57] S.-X. Zhang, X. Zhu, C. Yang, H. Wang, and X.-C. Yin, "Adaptive boundary proposal network for arbitrary shape text detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1305–1314.

[58] M. Liao, Z. Zou, Z. Wan, C. Yao, and X. Bai, "Real-time scene text detection with differentiable binarization and adaptive scale fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[59] 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.

[60] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6449–6458.

[61] M. Liao, P. Lyu, M. He, C. Yao, W. Wu, and X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[62] 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*, 2017, pp. 2550–2558.

Yadong Qu received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2020, where he is currently working toward the Ph.D. degree. 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 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.

Shancheng Fang received the Ph.D. degree in computer software and theory from the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, in 2020. He is currently a postdoctoral fellow at the School of Information Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include multimedia analysis and computer vision.

Yuxin Wang received the B.S. degree in XiDian University in 2018 and he is currently working toward the Ph.D. 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 (M'08–SM'13) 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.