IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 8, 2020, accepted August 17, 2020, date of publication August 24, 2020, date of current version September 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3018784

FDTA: Fully Convolutional Scene Text Detection With Text Attention

YONGCUN CAO, SHUAISEN MA, AND HAICHUAN PAN
School of Information Engineering, Minzu University of China, Beijing 100081, China
Corresponding author: Yongcun Cao (caoyongcun@126.com)

ABSTRACT Text detection is the premise and guarantee of text recognition. Multi-oriented text detection is the current research hotspot. Due to the variability in size, spatial layout, color and the arrangement direction of natural scene text, natural scene text detection is still very challenging. Therefore, this paper proposes a simple and fast multi-oriented text detection method. Our method first optimizes the regression branch by designing a diagonal adjustment factor to make the position regression more accurate, which increases F-score by 0.8. Secondly, we add an attention module to the model, which improves the accuracy of detecting small text regions and increases F-score by 1.2. Then, we introduce DR Loss to solve the problem of positive and negative sample imbalance, which increases F-score by 0.5. Finally, we conduct experimental verification and analysis on the ICDAR2015, MSRA-TD500 and ICDAR2013 datasets. The experimental results demonstrate that this method can significantly improve the precision and recall of scene text detection, and it has achieved competitive results compared with existing advanced methods. On the ICDAR 2015 dataset, the proposed method achieves an F-score of 0.849 at 9.9fps at 720p resolution. On the MSRA-TD500 dataset, the proposed method achieves an F-score of 0.772 at 720p resolution. On the ICDAR 2013 dataset, the proposed method achieves an F-score of 0.887 at 720p resolution.

INDEX TERMS Scene text detection, full convolution network, DR Loss, convolutional neural network.

I. INTRODUCTION

Scene text detection has important application value and research significance in real-time translation, image retrieval, scene analysis, geographic location and blind navigation. At present, there still needs some improvement in scene text detection. Text detection methods mainly include traditional text detection methods and text detection methods based on deep learning.

Traditional text detection methods mainly use machine learning methods for classification. These machine learning methods mainly include neural networks, SVM, K-Means and so on. The neural network is a dynamic system with a topological structure of directed graph, which processes information by responding to continuous or intermittent inputs. The neural network has many applications in the latest research [1]–[6]. SVM (Support Vector Machine) is a kind of generalized linear classifier that classifies data binary by supervised learning. K-Means clustering algorithm is an iterative clustering algorithm. Traditional text detection methods can be divided into sliding window methods and connected region methods. The sliding window-based methods [7], [8] mainly use sliding windows of different scales to search the whole image and extract the features in the window. Common text detection methods based on connected regions include Maximum Stable Extremum Regions (MSER), extremum region method and Stroke Width Transformation (SWT) [9], [10]. Traditional text detection methods usually include multiple steps: generating candidate regions, filtering candidate regions, constructing text lines and verifying text lines. Every module needs to be well designed in order to achieve good performance. These methods can achieve good performance in simple scenes. However, in complex natural scenes, such as uneven illumination or partial occlusion, the traditional machine learning method still cannot achieve an ideal detection performance.

In recent years, compared with traditional text detection methods, the text detection methods based on deep learning have made a great breakthrough in performance. Among the text detection methods based on deep learning, the regression-based methods and the segmentation-based methods are the most widely used. The regression-based methods are mainly based on the improvement of text characteristics in the object detection framework, such

The associate editor coordinating the review of this manuscript and approving it for publication was Shiping Wen.

as SSD [11], Faster-RCNN [12] and YOLO [13]. These methods obtain the detection result by regressing the shape of a horizontal rectangular, rotating rectangular and quadrilateral. The segmentation-based methods usually use the idea of semantic segmentation, and divide text pixels into different instances, and obtain text pixel-level positioning results through some post-processing methods. However, post-processing is usually more complex and consumes a lot of computing resources. Both regression-based methods and segmentation-based methods have their limitations.

Because of the above problems, this paper proposes a simple and efficient multi-oriented text detection method, which is robust to changes in the text scale. Because FCOS [14] has good performance in object detection based on anchor-free prediction, we propose a multi-oriented scene text detection method based on the attention mechanism.

Our main contributions can be summarized as follows:

1)A diagonal adjustment factor is designed to regress the loss function, making the position regression more accurate.

2)The text attention module is added, which makes the text pay more attention to useful information and restrain useless information.

3)Using DR Loss improves the imbalance between positive and negative samples, which further improves network performance.

4)Our method achieves competitive results in terms of speed and accuracy on some standard text detection benchmarks.

II. RELATED WORKS

In recent years, with the rapid development of deep learning, scene text detection has made great progress. This section focuses on the work most relevant to the method proposed in this paper.

The EAST [15] method combines multi-scale feature maps for dense pixel-by-pixel prediction. EAST also uses the rotating box (RBOX) and arbitrary quadrilateral (QUAD) for position prediction. However, due to the lack of a receptive field, EAST cannot detect long text effectively. The TextBoxes++ [16] method modifies the anchor of SSD and makes the performance better. TextBoxes++ also combines text recognition to improve accuracy. The RRD [17] method distinguishes the classification and regression of text detection to better detect long texts. The FTSN [18] method is an end-to-end trainable multi-oriented text detection based on instance-aware segmentation. The Inceptext [19] method proposes a deformable PSROI pooling module to process multi-oriented text. The Masktext Spotter [20] method proposes a mask text detection, which can detect and recognize the text of any shape. In addition, some methods regard text detection as semantic segmentation of text regions. In [21], Shi et. al predicts the connection between the text instance fragments and then connects them into text instances. In [22], Deng et. al fuses multi-layer depth features to improve detection accuracy. The Corner [23] method integrates the detection and segmentation methods into a comprehensive score. The TextSnake [24] method uses orderly overlapping disks to represent text lines of arbitrary shape. FCOS is a full convolutional one-stage object detection algorithm, which solves the object detection problem by pixel prediction. The solutions of anchor free and proposal free are realized, and the idea of Center-ness is put forward. At the same time, the, recall approaches or even exceeds many anchor-based object detection algorithms. By removing the predefined anchor, FCOS completely avoids the complex operation of anchor and saves memory occupation during training.

Different from the above methods, this paper proposes a multi-oriented scene text detection method based on the FCOS. Based on the FCOS anchor-free prediction, the regression method is designed for multi-direction text. On this basis, the diagonal adjustment factor is designed, which makes the position regression more accurate. In addition, when extracting features of texts, a text attention module is added to improve important features and reduce the interference of irrelevant information. Compared with directly using the object detection models, this method can ensure that small text areas are not lost, and ensure the integrity of narrow text areas detection.

III. THE PROPOSED METHOD

Our method is based on the FCOS network, which uses pixel-by-pixel prediction to detect text. It directly predicts the distance between the 4 points of the spatial point to the object on the feature map. This method does not need the anchor and avoids the complicated calculation related to the anchor. Because we use the high-resolution feature map, our method can detect very small texts. In this section, we will show our method in detail.

A. NETWORK STRUCTURE

Due to the good performance of FCOS, we design a text detection network based on FCOS. Figure 1 shows our network structure, which mainly includes three parts: feature extraction network, feature-merging branch and output layer. The feature extraction network uses Darknet53 [25] as the basic network for extracting text features. Darknet53 has five levels of feature maps. As shown in Figure 1, this paper mainly uses four levels of a feature layer, whose sizes are respectively 1/32, 1/16, 1/8 and 1/4 of the input image. The feature map extracted from Conv Stage5 is sampled and processed to enlarge the size to 1/16 of the input image. Then, the extracted feature map is combined with Conv Stage4. After merging the feature map, the 1×1 convolution operation is used to fuse the feature map and reduce the number of channels. Through four Conv blocks, the final output of the whole fusion branch is obtained as the input of the attention module. The attention module weights the extracted features, highlighting important feature information and weakening the irrelevant information. The content of the module will be introduced in detail in the third part. Finally, the output layer contains three branches, which are 1 channel text score

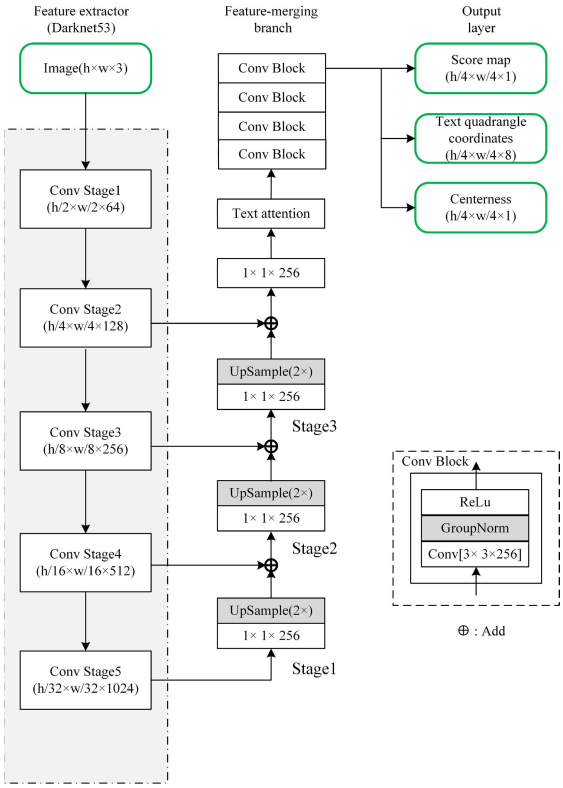

FIGURE 1. Structure of our text detection.

branch, 8 channel location regression branch and 1 channel center branch.

In our approach, each target is expressed as $[x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4]$.

$$\begin{aligned} x &= \left\lfloor \frac{s}{2} \right\rfloor + xs, \\ y &= \left\lfloor \frac{s}{2} \right\rfloor + ys \end{aligned} \tag{1}$$

where [x, y] is the location on the image, and [xs, ys] is the location on the feature, and s represents the step size in the feature map. The regression branch predicts the object offset and outputs the 8D vector $[x_{lt}, y_{lt}, x_{rt}, y_{rt}, x_{rb}, y_{rb}, x_{lb}, y_{lb}]$. It represents every position in the feature map and corresponds to the position on the image, which is calculated as follows:

$$\begin{aligned} x_{lt} &= x - x_1, \quad y_{lt} = y - y_1, \\ x_{rt} &= x_2 - x, \quad y_{rt} = y - y_2, \\ x_{rb} &= x_3 - x, \quad y_{rb} = y_3 - y, \\ x_{lb} &= x - x_4, \quad y_{lb} = y_4 - y. \end{aligned} \tag{2}$$

In addition, in order to speed up convergence, the regression branch is divided into strides by train and inference.

FCOS introduces Center-ness to suppress the low-quality detection box without introducing any super parameters, which proves its effectiveness in object detection. However, the vector of our regression is eight-dimensional. Therefore, we design the eight-dimensional vector of the center regression corresponding to the regression vector $[x_{lt}, y_{lt}, x_{rt}, y_{rt}, x_{rb}, y_{rb}, x_{lb}, y_{lb}]$. We calculate the four-dimensional vectors $[lt, rt, rb, lb]$, representing the distance from the center point to the four vertices. Each vector can be written as:

$$\begin{aligned} lt &= \sqrt{(x_{lt} + y_{lt})^2}, \quad rt = \sqrt{(x_{rt} + y_{rt})^2}, \\ rb &= \sqrt{(x_{rb} + y_{rb})^2}, \quad lb = \sqrt{(x_{lb} + y_{lb})^2}. \end{aligned} \tag{3}$$

In this respect, we propose quadrilateral centerness:

$$centerness = \sqrt{\frac{\min(lt, rb)}{\max(lt, rb)} \times \frac{\min(lb, rt)}{\max(lb, rt)}}, \tag{4}$$

The square root is used to slow down the attenuation of centerness. The range of centerness is from 0 to 1, so training is conducted through binary cross entropy (BCE) loss. The loss is added to the loss function formula [5]. When testing, the final score is calculated by multiplying the predicted centerness by the corresponding class score. Therefore, centerness can reduce the score weight of the bounding box far from the center of the object.

B. LOSS FUNCTION

This section describes the loss function of this model. The overall loss function of the model is as follows:

$$L = L_{cls} + \frac{\lambda}{N_{pos}} L_{reg} + \frac{\omega}{N_{pos}} L_{center}, \tag{5}$$

where L_{cls} is predicting score loss. L_{reg} represents regression quadrilateral loss. L_{center} indicates centerness loss. N_{pos} represents the number of positive samples in ground truth. λ and ω are balance factors. λ and ω set to 1.

Class imbalance can reduce the performance of the class prediction of text detection models. Each image contains a large number of candidate boxes, usually 10 K candidate boxes. However, the bounding box of a real image may only have one, or even none, which leads to the imbalance between positive samples and negative samples. Meanwhile, negative samples consume a lot of computing resources.

The deep learning models solve this problem by data augmentation or hard negative mining in the training process. However, these practices introduce extra steps or introduce a non-differential stage in the whole detection process. Different from these methods, we introduce DR Loss to solve this problem. Because DR Loss can transform classification problems into sorting problems, the problem of imbalance between positive and negative samples is improved. L_{cls} expressed as follows:

$$\ell_{logistic}(z) = \frac{1}{L} \log(1 + \exp(Lz)), \tag{6}$$

$$L_{cls} = \sum_{i}^{N} \ell_{\text{logistic}} \left(P_{i,-} - P_{i,+} + \gamma \right), \tag{7}$$

where P_- and P_+ are the distributions of positive and negative samples. L is the approximate error of the control function. γ ensures that the positive and negative samples can be

separated. In this experiment, L is set to 6 and γ to 2. More details of DR Loss can be found in the paper [26].

In terms of regression loss, we use smoothed-L1 loss [7]. Q is an ordered set of all coordinate values, which can be written as follows:

$$C_Q = [x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4]. \tag{8}$$

and then the loss is calculated as:

$$L_{loc} = L_{QUAD}\left(\hat{Q}, Q^*\right) = \sum_{\substack{c_i \in C_Q \\ \tilde{c}_i \in C_{\tilde{Q}}}} smoothed_{L1}\left(c_i - \tilde{c}_i\right). \tag{9}$$

However, in natural scenes, the aspect ratio is extremely changeable. The smoothed-L1 loss does not consider the correlation of the coordinate points, which leads to inaccurate boundary prediction. In object detection, IOU Loss [27] is usually used to solve this problem. However, it usually requests the rectangle box from IOU in object detection. It is more time-consuming to calculate IOU for the quadrilateral. Therefore, inspired by CIOU Loss [28], we consider the diagonal proportion of the predicted object to fit the diagonal proportion of the ground truth. The diagonal adjustment factor is introduced, which significantly improves the accuracy of boundary prediction. More details can be found in the ablation study part (shown in Table 4). Diagonal adjustment factors are as follows:

$$\alpha = \frac{4}{\pi^2}\left(arctan\frac{d_1^{gt}}{d_2^{gt}} - arctan\frac{d_1}{d_2}\right), \tag{10}$$

$$d_1 = \sqrt{(x_3 - x_1)^2 + (y_3 - y_1)^2},$$

$$d_2 = \sqrt{(x_4 - x_2)^2 + (y_4 - y_2)^2}, \tag{11}$$

where d_1, d_2 represent the length of the diagonal of the predicted box, and d_1^{gt}, d_2^{gt} represent the ground truth diagonal length. The formula is equivalent to adjusting the aspect ratio, making the prediction more robust. Finally, the loss of regression can be written as:

$$L_{reg} = (1 + \alpha) \times L_{loc}. \tag{12}$$

According to FCOS [14], the centerness loss is constructed, and the standard binary cross entropy loss are extended to the centerness loss. The purpose of centerness loss is to encourage the network to choose a regression point close to the object center. Besides, the centerness also affects the confidence level of the predicted object. The loss can be written as:

$$L_{center} = BCE\left(P_{centerness}; G_{centerness}\right). \tag{13}$$

C. ATTENTION MODULE

The attention mechanism is derived from the research of human visual image information, mainly for the rational use of limited resources to represent the whole thing itself. The attention mechanism can highlight important features, and reduce the interference of irrelevant information on detection results. Therefore, we introduce the attention mechanism to improve the accuracy of prediction. The Dual Attention Network [29] adaptively integrates the local features and global dependency to capture the global feature, which enhances the feature representation of scene segmentation. We separate the channel attention module to use in the text detection model, which improves the prediction accuracy without increasing the calculation and reducing the inference speed.

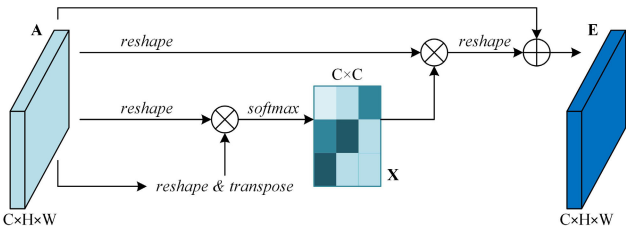

FIGURE 2. The text attention module [29].

As shown in Figure 2, the channel attention map $X \in R^{C\times C}$ is directly calculated from the initial $A \in R^{C\times H\times W}$. And the shape of A to $R^{C\times N}$ changes. Then, matrix multiplication is performed between A and A transpose. Finally, the channel attention map is obtained through a softmax layer. It can be written as:

$$x_{ji} = \frac{\exp\left(A_i \cdot A_j\right)}{\sum_{i=1}^{C} \exp\left(A_i \cdot A_j\right)}, \tag{14}$$

where x_{ij} indicates the influence of i^{th} channel on j^{th} channel. Besides, the matrix multiplication between X and A transpose is made.

$$E_j = \beta \sum_{i=1}^{C}\left(x_{ji}A_i\right) + A_j, \tag{15}$$

where β is initialized to 0, and then gradually learns. Formula [14] shows that the final feature of each channel is the weighted sum of all original features and channel features.

As shown in Figure 1, our channel attention module is represented as text attention, which processes the output of feature fusion and enhances the ability of feature expression.

IV. EXPERIMENTS

In order to evaluate of our method, we test it on ICDAR2013 [30], ICDAR2015 [31] and MSRATD-500 [32]. We compare our method with the latest text detection methods, and verify that our method can achieve the most advanced level. Moreover, we also carry out ablation experiments to study the effects of text channel attention, diagonal adjustment factors and DR Loss on the overall model performance.

A. BENCHMARK DATASETS

ICDAR2015 [31] is the open source dataset of the ICDAR competition in 2015. The images in the dataset are captured by Google Glass. The background environment is all random natural scenes. The ICDAR2015 dataset contains 1000 images for training and 500 images for testing.

Compared with other image databases, the image quality of ICDAR2015 dataset is not high. The size and direction of the text in the image are different. The text annotations in the image are given in the form of the word unit.

MSRA-TD500 [32] contains 500 pictures, 300 for training and 200 for testing. All pictures are indoor (office, shopping mall) or outdoor (street) scenes captured by portable cameras. The dataset contains not only English text but also Chinese text, with the text line as the unit.

ICDAR2013 [30] contains 229 training images and 233 testing images. The text instance is almost horizontal.

The evaluation index of text detection depends on Precision (P), Recall (R) and F-score (F), which are defined as:

$$P = \frac{TP}{TP + FP},$$
$$R = \frac{TP}{TP + FN},$$
$$F = 2 \times \frac{P \times R}{P + R}, \tag{16}$$

where TP, FP and FN are the correct detections, the wrong detections and the number of missing detections respectively. In addition, frames per second (FPS) are usually used to measure the performance of text detection algorithms.

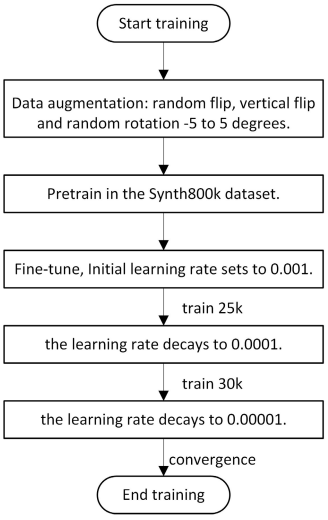

FIGURE 3. Training flowchart. Firstly, all training images are subjected to data augmentation, including random horizontal flip, vertical flip and random rotation −5 to 5 degrees. Secondly, 100k images are randomly selected from the Synth 800k dataset for pre-training. Then, the model is fine-tuned with real dataset such as ICDAR2015. The initial learning rate is 0.001. When iterating 25k, the learning rate decays to 0.0001. When iterating 30k, the learning rate decays to 0.00001. Finally, the training is complete until convergence.

B. IMPLEMENTATION DETAILS

We use the model trained on ImageNet dataset [33] as our pre-trained model. As shown in Figure 3, the training process includes two steps: we first randomly extract 100k pictures from the Synth800k [34] dataset to train 10 epochs, and then fine-tune the model with real dataset until convergence.

Our data augmentation methods include random horizontal flip, vertical flip and random rotation −5 to 5 degrees. Our network optimizes the model through the SGD and momentum method, making the weight decay coefficient of the network set to 5×10^{-4}. The momentum sets to 0.9 and the batch size sets to 16. The initial learning rate sets to 0.001 and the learning rate is reduced by 10 times when 25K and 30K are iterated. We adjust the size of the input image to make the width equal to 1280 pixels, and the height equals 720 pixels. All experiments are implemented in Python, using PyTorch 1.3. Our model runs in Ubuntu 16.04 system and Nvidia Tesla V100.

TABLE 1. Results on ICDAR2015. "P", "R", "F" represent "Precision", "Recall" and "F-score" respectively. The method with "∗" means multi-scale testing. Mask text means mask textspotter.

Algorithm	R (%)	P (%)	F (%)	FPS	Published
SegLink [21]	76.8	73.1	75.0	20.6	2017
PixelLink [22]	82.0	85.5	83.7	3	2018
Corner [23]	94.1	70.7	80.7	3.6	2018
Corner*	89.5	79.7	84.3	1	2018
IncepText [19]	80.6	90.5	85.3	3.7	2018
EAST [15]	73.4	83.5	78.2	13	2017
EAST*	78.3	83.2	80.7	6.5	2017
TextBoxes++ [16]	76.7	87.2	81.7	11.6	2018
TextBoxes++*	78.5	87.8	82.9	-	2018
RRD [17]	79.0	85.6	82.2	6.5	2018
FSTN [18]	80.0	88.6	84.1	2.5	2018
TextSnake [24]	80.4	84.9	82.9	1.1	2018
Mask Text [20]	81.2	85.8	83.4	4.8	2018
Richard et al. [35]	83.1	85.4	84.2	3.5	2020
DSRN [36]	79.6	83.2	81.4	8.8	2019
Ours	**81.2**	**89.0**	**84.9**	**9.9**	-

C. COMPARISON WITH OTHER METHODS

Table 1 shows the qualitative comparison on ICDAR2015. Our method achieves the highest F-score 84.9% and our speed is 9.9 FPS without the multi-scale test. Compared with EAST*, our method increases 2.9%, 5.8% and 4.2% respectively on the three indexes of R, P and F. This is because the receptive field of EAST is very small, which leads to the poor detection of long texts. Our network expands the receptive field, making the long text of detection more accurate. It also proves the validity of the diagonal regulator module to the regression branch. Compared with the latest method proposed by Richard [35], our method increases 0.7% in F-score and the inference speed is 2.8 times faster. This indicates that our method is still competitive with the latest text detection methods.

We also evaluate our method on MSRA-TD500. Because the object size of MSRA-TD500 is large and the text is complex, the effect of the existing text detection method on MSRA-TD500 is worse than that of ICDAR2015. As shown in Table 2, Liang proposes a text detection method based on

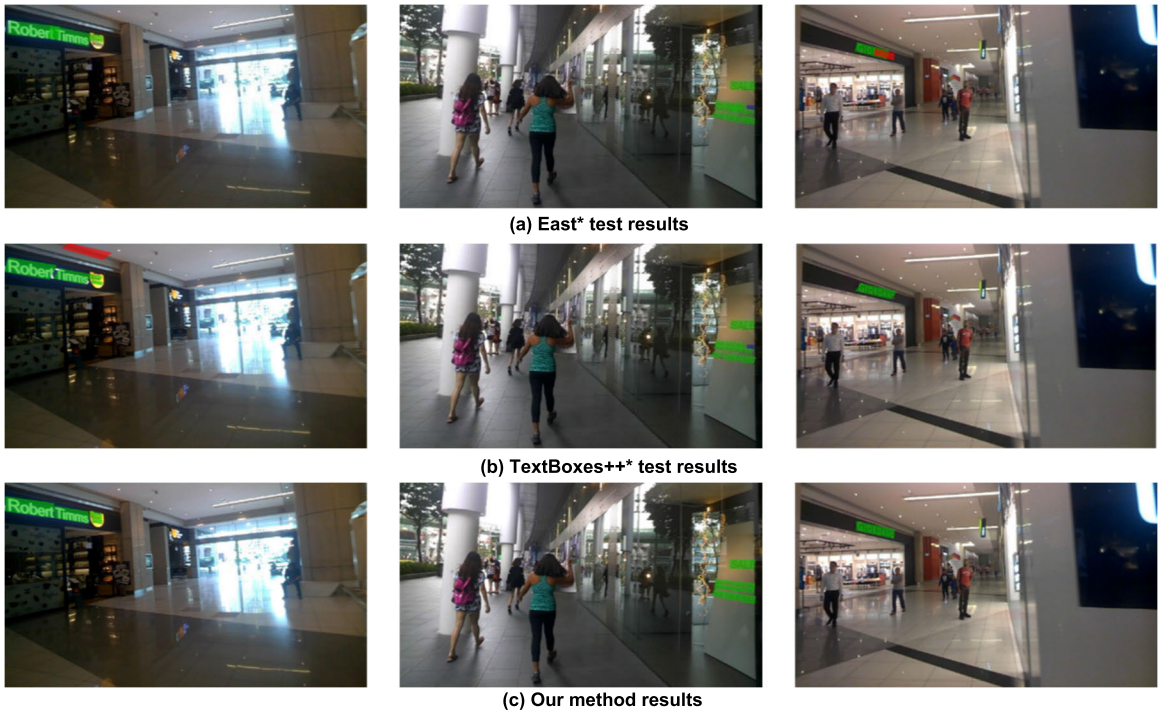

FIGURE 4. Comparison of text detection results on ICDAR2015. Green box area indicates correct detection; red box area indicates error detection; blue box area indicates missing detection.

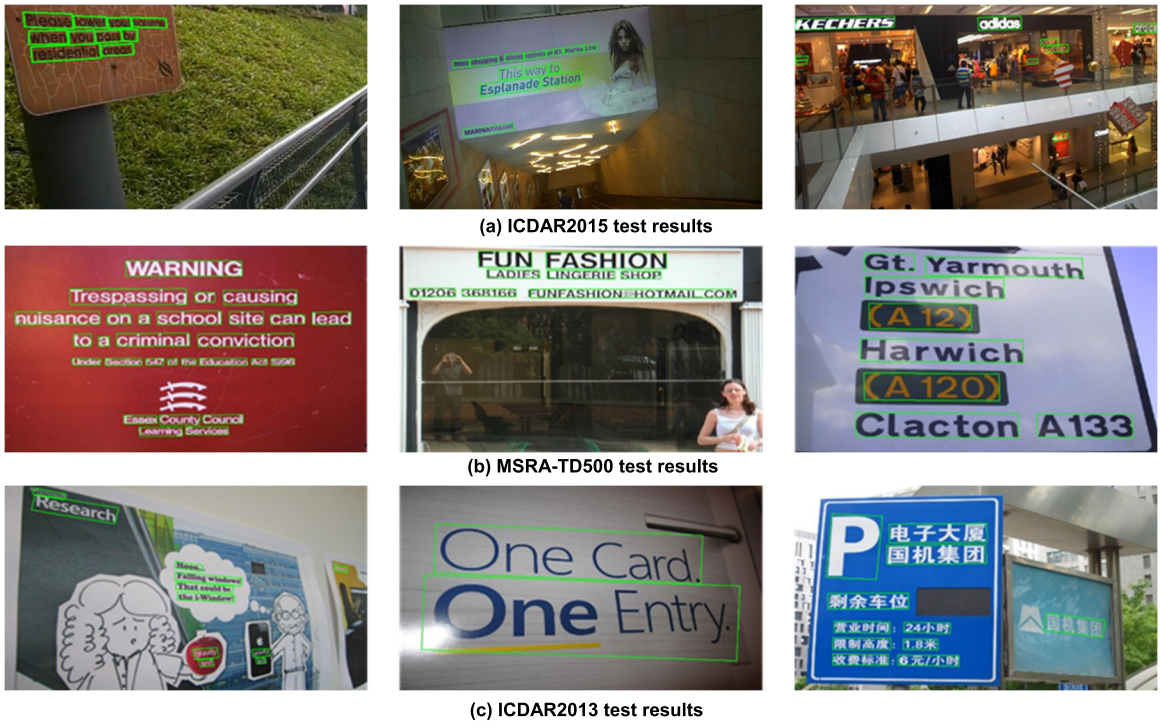

FIGURE 5. Some examples of our method in ICDAR2015, MSRA-TD500 and ICDAR2013.

machine learning, which is 7.2% lower than our method in F-score. Zhang *et al.* is an advanced multi-oriented text detection method published before. Compared with this method, our method exceeds 4.2%, 1.2% and 3.2% in three indexes respectively. The F-score of our method is 1.2% higher than that of the EAST method. The comparison result is shown

FIGURE 6. Some examples of large-scale texts.

TABLE 2. Results on MSRA-TD500. "P", "R", "F" represent "Precision", "Recall" and "F-score" respectively.

Algorithm	R (%)	P (%)	F (%)	Published
Liang et al. [37]	66.0	74.0	70.0	2015
Zhang et al. [38]	67	83	74	2016
Yao et al. [39]	75.3	76.5	75.9	2016
SegLink [21]	70	86	77	2017
EAST [15]	67.3	87.2	76.0	2017
He et al. [40]	70	77	74	2016
Ours	**71.2**	**84.2**	**77.2**	-

in Table 3. We also conduct experiments on ICDAR2013, which is a popular horizontal text dataset. Our method achieves the best F-score on ICDAR2013. FASText is a text detection method based on machine learning, which is 11.9% lower than our method in F-score. Compared with SegLink, our method exceeds 1.6%, 5.5%, 3.4% on three indexes respectively.

TABLE 3. Results on ICDAR2013. "P", "R", "F" represent "Precision", "Recall" and "F-score" respectively.

Algorithm	R (%)	P (%)	F (%)	Published
Liang et al. [37]	68.0	76.0	72.0	2015
FASText [41]	69.3	84.0	76.8	2015
SegLink [21]	83.0	87.7	85.3	2017
PixelLink [22]	83.6	86.4	84.5	2018
Corner [23]	79.4	93.3	85.8	2018
TextBoxes ++ [16]	74.0	86.0	80.0	2018
Ours	**84.6**	**93.2**	**88.7**	-

Figure 4 and 5 show some detection examples of our method. In Figure 4, we compare our method with EAST and TextBoxes++. From the left image, we can see that when TextBoxes++ misses the ceiling as a text area. Our method will not be confused. In the middle image, EAST and TextBoxes++ omit a small text area. Our method can still be detected. In addition, from the right image, we can see that EAST detects long text breakage and TextBoxes++ cannot completely encircling text information. We have a good effect on the detection of long text, and the detection box is more accurate. In Figure 5, we show some samples of our method on ICDAR2015, MSRA-TD500 and ICDAR2013. Some examples of detection on MSRA-TD500 and ICDAR2013 show that our method not only detect English text, but also detect Chinese text with different fonts. Therefore, our method is robust to different font and irregular structure.

Figure 6 shows some detection examples of large-scale texts. From the left image, our detector can completely detect ordinary large-scale text. From the middle image, our detector cannot completely detect the edge of the long large-scale text. From the right image, the detection boundary is incomplete for very large-scale texts. There are two reasons for the loss of large-scale text boundaries. On the one hand, it is the limitations of the receptive field of the network. On the other hand, in the process of training, there are few super large-scale text samples and the learned features are incomplete.

D. ABLATION STUDY

In order to directly observe the function of each component in the model, the ablation experiment is conducted in this section. Because ICDAR2015 is the most influential and difficult, the result of the dataset can better reflect the practicability of the method. Therefore, the whole experiment is carried out on the dataset to study the influence of three components on multi-oriented text detection: text attention mechanism, diagonal factor and DR Loss. All the experiment is the same except for the control variables in this section, and the experimental results are shown in Table 4.

TABLE 4. Ablation experiments for different components on ICDAR2015. "TA" means "Text attention module" and "DF" means "Diagonal adjustment factor" and "DR" means "DR Loss". "P", "R", "F" represent "Precision", "Recall" and "F-score" respectively.

	TA	DF	DR	R (%)	P (%)	F (%)
baseline				80.8	85.0	82.8
Ours	√			83.0	85.0	84.0
		√		78.8	89.1	83.6
			√	79.0	88.2	83.3
		√	√	79.1	89.3	83.9
	√	√		79.7	90.4	84.7
	√		√	80.1	89.4	84.5
	√	√	√	81.2	89.0	84.9

We choose the Darknet53 model as the baseline of this experiment. From the Table 4, we can see that our text attention module can help the model to increase the F-score by 1.2%. The experimental results show that the attention module can be used to learn more important features and

enhance the representation ability of the model. Second, our diagonal adjustment factor makes the F-score increase by 0.8% again. In addition, DR Loss is introduced as the classification loss, which increases F-score by 0.5%. This group of experiment shows that the positive and negative sample imbalance affects the network performance, and sDR Loss improves this problem and the network performance. When using the combination of diagonal adjustment factor and DR Loss, F-score is improved by 1.1%. Finally, when using these three components, compared with the baseline, F-score is improved by 2.1%.

V. CONCLUSION

In this paper, a multi-oriented text detection method combined with an attention module is proposed. This method is based on the FCOS network and uses pixel by pixel prediction to detect text. It directly predicts the distance between these four boundaries of the spatial point and the object on the feature map. In order to make the position regression more accurate, the diagonal adjustment factor for the regression loss function is designed, which increases F-score by 0.8. In addition, the text attention module is added, which pays more attention to useful information and increases F-score by 1.2. Then, aiming at the problem of sample imbalance in text detection, DR Loss is introduced to enhance the detection performance of the network and increase F-score by 0.5. Finally, we perform experimental comparison and model analysis on scene text detection datasets ICDAR2015, MSRA-TD500 and ICDAR2013. On the ICDAR 2015 dataset, the proposed method achieves an F-score of 0.849 at 9.9fps at 720p resolution. The results show that the method has achieved an advanced level. At the same time, we have verified the function of the components in this paper through ablation experiments. However, our detector still has room for improvement. Firstly, for very large-large text, the boundaries our detector detects are incomplete. We plan to further improve this problem. In addition, our detector is currently designed for multi- oriented text, which is not suitable for curve text. We plan to further improve our network processing and output to achieve this.

REFERENCES

[1] J. Lu, J. Xuan, G. Zhang, and X. Luo, "Structural property-aware multilayer network embedding for latent factor analysis," *Pattern Recognit.*, vol. 76, pp. 228–241, Apr. 2018.

[2] B. Sun, S. Wen, S. Wang, T. Huang, Y. Chen, and P. Li, "Quantized synchronization of memristive neural networks with time-varying delays via super-twisting algorithm," *Neurocomputing*, vol. 380, pp. 133–140, Mar. 2020.

[3] B. Sun, Y. Cao, Z. Guo, Z. Yan, and S. Wen, "Synchronization of discrete-time recurrent neural networks with time-varying delays via quantized sliding mode control," *Appl. Math. Comput.*, vol. 375, Jun. 2020, Art. no. 125093.

[4] S. Wang, Y. Cao, T. Huang, Y. Chen, and S. Wen, "Event-triggered distributed control for synchronization of multiple memristive neural networks under cyber-physical attacks," *Inf. Sci.*, vol. 518, pp. 361–375, May 2020.

[5] W.-J. Niu, Z.-K. Feng, Y.-B. Chen, H.-R. Zhang, and C.-T. Cheng, "Annual streamflow time series prediction using extreme learning machine based on gravitational search algorithm and variational mode decomposition," *J. Hydrol. Eng.*, vol. 25, no. 5, May 2020, Art. no. 04020008.

[6] Z.-K. Feng, W.-J. Niu, R. Zhang, S. Wang, and C.-T. Cheng, "Operation rule derivation of hydropower reservoir by k-means clustering method and extreme learning machine based on particle swarm optimization," *J. Hydrol.*, vol. 576, pp. 229–238, Sep. 2019.

[7] K. In Kim, K. Jung, and J. Hyung Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1631–1639, Dec. 2003.

[8] Y.-F. Pan, X. Hou, and C.-L. Liu, "A robust system to detect and localize texts in natural scene images," in *Proc. 8th IAPR Int. Workshop Document Anal. Syst.*, Sep. 2008, pp. 35–42.

[9] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, Sep. 2004.

[10] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, Nov. 2010, pp. 770–783.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2016, pp. 21–37.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[14] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

[15] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5551–5560.

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

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

[18] Y. Dai, Z. Huang, Y. Gao, Y. Xu, K. Chen, J. Guo, and W. Qiu, "Fused text segmentation networks for multi-oriented scene text detection," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3604–3609.

[19] Q. Yang, M. Cheng, W. Zhou, Y. Chen, M. Qiu, W. Lin, and W. Chu, "IncepText: A new inception-text module with deformable PSROI pooling for multi-oriented scene text detection," 2018, *arXiv:1805.01167*. [Online]. Available: http://arxiv.org/abs/1805.01167

[20] 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. Vis. (ECCV)*, 2018, pp. 67–83.

[21] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2550–2558.

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

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

[24] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "Textsnake: A flexible representation for detecting text of arbitrary shapes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 20–36.

[25] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[26] Q. Qian, L. Chen, H. Li, and R. Jin, "DR loss: Improving object detection by distributional ranking," 2019, *arXiv:1907.10156*. [Online]. Available: http://arxiv.org/abs/1907.10156

[27] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. ACM Multimedia Conf.*, 2016, pp. 516–520.

[28] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," 2019, *arXiv:1911.08287*. [Online]. Available: http://arxiv.org/abs/1911.08287

[29] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[30] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. I. Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras, "ICDAR 2013 robust reading competition," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 1484–1493.
[31] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny, "ICDAR 2015 competition on robust reading," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1156–1160.
[32] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1083–1090.
[33] A. Krizhevsky and I. G. E. S. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[34] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2315–2324.
[35] E. Richardson, Y. Azar, O. Avioz, N. Geron, T. Ronen, Z. Avraham, and S. Shapiro, "It's all about the Scale–Efficient text detection using adaptive scaling," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1844–1853.
[36] Y. Wang, H. Xie, Z. Fu, and Y. Zhang, "DSRN: A deep scale relationship network for scene text detection," in *Proc. Twenty-Eighth Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 947–953.
[37] G. Liang, P. Shivakumara, T. Lu, and C. L. Tan, "Multi-spectral fusion based approach for arbitrarily oriented scene text detection in video images," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4488–4501, Nov. 2015.
[38] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4159–4167.
[39] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao, "Scene text detection via holistic, multi-channel prediction," 2016, *arXiv:1606.09002*. [Online]. Available: http://arxiv.org/abs/1606.09002
[40] 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.
[41] M. Buta, L. Neumann, and J. Matas, "FASText: Efficient unconstrained scene text detector," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1206–1214.

YONGCUN CAO received the B.S. degree, in 1986. He is currently a Professor with the School of Information Engineering, Minzu University of China, Beijing, China. His current research interests include big data, parallel algorithm, and intelligent systems.

SHUAISEN MA received the B.S. degree from the Tianjin University of Science and Technology, China, in 2018. He is currently pursuing the master's degree with the School of Information Engineering, Minzu University of China, Beijing, China. His research interests include deep learning and scene text detection.

HAICHUAN PAN received the B.S. degree from the Luoyang Institute of Science and Technology, China, in 2019. He is currently pursuing the master's degree with the School of Information Engineering, Minzu University of China, Beijing, China. His research interests include machine learning and swarm intelligence.