IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 20, 2019, accepted December 8, 2019, date of publication December 12, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2959018

IoU-Related Arbitrary Shape Text Scoring Detector

FAGUI LIU, DIAN GU, AND CHENG CHEN
School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China

Corresponding author: Dian Gu (546780523@qq.com)

This work was supported in part by the Engineering and Technology Research Center of Guangdong Province for Logistics Supply Chain and Internet of Things under Project GDDST[2016]176, in part by the Key Laboratory of Cloud Computing for Super-integration Cloud Computing in Guangdong Province under Project 610245048129, in part by the Engineering and Technology Research Center of Guangdong Province for Big Data Intelligent Processing under Project GDDST[2013]1513-1-11, in part by the "Made in China 2025" Industrial Development Fund Project of Guangzhou under Project x2jsD8183470, and in part by the Major Program of Guangdong Basic and Applied Research under Project 2019B030302002.

ABSTRACT As a medium of information transmission, text is widely existed in natural scenes, showing diversity in orientation, scale, font, color and shape. Accurate detection of scene text is a prerequisite for subsequent recognition. Though many previous methods have worked well on horizontal and multi-oriented text detection datasets, detecting arbitrary shape scene text still remains as a challenging problem. To solve the problem, this paper proposes an arbitrary shape scene text detection method. Based on Mask R-CNN, our method replaces the original ℓ_1-smooth loss with the proposed IoU-related loss and adds a text scoring branch to align the confidence score with the text mask IoU to make the model highly relevant to IoU, achieving the goal of improving detection performance by improving IoU directly in a simple but effective way. The proposed method is evaluated on four public datasets: CTW-1500, Total-Text, ICDAR2015 and ICDAR2017-RCTW. For curved text detection datasets CTW-1500 and Total-Text, we have reached 79.2% and 81.1% H-mean respectively, showing that the proposed method has achieved competitive performance in arbitrary scene text detection.

INDEX TERMS Scene text detection, arbitrary shape, computer vision, intersection-over-union, semantic segmentation.

I. INTRODUCTION

Scene text detection is an important branch of object detection, also an attracting task in computer vision, as it can be widely used in many applications as image search, automatic transmission, blind person assistance, real-time translation and so on [1], [2]. Accurate text localization is a crucial premise for subsequent recognition task. However, there are a series of challenges when detecting text in the wild. Firstly, scene text exhibits much higher diversity and variability. For instance, text in the wild can be multi-lingual with drastic scale changes, and has different fonts, colors, shapes and orientations, which adds difficulty to text detection and recognition. Secondly, scene text detection occasionally suffers from false positive error due to the similarity between text instance and background texture, resulting in a decrease in detection performance. For example, patterns such as fences, tree leaves, curtain texture or bricks make it hard to distinguish from text instance. Thirdly, imperfect image quality is also a distracting factor. Unlike scanned scripts in documents, manually captured scene text can be distorted and out of focus, which affects the detection effect seriously. Moreover, due to different lighting conditions and shooting angles, manual collection inevitably introduces noise and occlusion.

Early studies like connected component(CC) based [3]–[5] or sliding window based [6]–[8] approaches used manually designed features and traditional classifiers for text detection, but they were not well adapted to the diversity and variability of scene text. Benefitted from the introduction of deep learning, the detection performance has been significantly improved. With the development of deep learning based approaches, the research focus of scene text detection has shifted from horizontal scene texts to more challenging tasks such as incidental multi-oriented scene texts [9], curved or

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

arbitrary shape scene texts [10], [11] and multi-lingual scene texts [12].

Some Faster R-CNN based methods [13]–[15] achieve competitive performance on multi-oriented datasets, but are not ideal when dealing with arbitrary shape text detection challenge. Although they can detect most of the text instances, there remains needless overlap and redundant background noise, which are very harmful to detecting performance. Moreover, for text detection task, the ultimate goal of improving evaluation scores is to improve the Intersection-over-Union(IoU) between proposals and corresponding ground truth. These methods abovementioned focus more on changing the scale, aspect ratio, rotation angle and shape of anchor, or changing the network framework, or introducing new modules to achieve better scores, while there still remains two problems. The first on the list is that, there is not a strong corelation between minimizing ℓ_1-smooth loss and improving the IoU value between the proposed and the ground truth bounding box. The second on the list is that there is not a strong corelation between the classification confidence score and the text maskIoU value, so proposed text mask with high IoU value but low confidence score would be filtered by non-maximum suppression(NMS).

To overcome the shortcomings mentioned above, we propose an arbitrary shape scene text detection method based on generalized object detection structure Mask R-CNN [16]. Without too many changes to the model structure, we replace the loss function of box regression branch in RPN and box regression branch in Mask RCNN heads to make a strong correlation between loss and IoU, and add text scoring branch, which predicts the IoU of text mask provided by mask branch. The predicted IoU represents the quality of the mask. We use this predicted IoU to multiply the classification confidence score to get the final confidence score, so as to align it with the mask quality.

Our contributions are summarized as follows:

1) We propose an IoU related scene text detector for arbitrary shape text detection task. A text scoring branch is added as an ROI head of Mask R-CNN to align the confidence score with the text mask quality.
2) We propose a generalized completeness-aware IoU related loss function in replace of ℓ_1-smooth loss in regression branches.
3) We conduct experiments on several public arbitrary shape, multi-oriented and multi-lingual scene text datasets, including CTW-1500, Total-Text, ICDAR2015 and ICDAR2017-RCTW, to prove the superiority of our method over previous ones.

The rest of the paper is organized as follows. In Section II, we review some of the prior proposed methods in scene text detection; Section III introduces the scene text detection framework and the loss function we propose; and in Section IV, experimental details and detection results on three public datasets are shown. Section V is our final conclusion and our prospective future work.

II. RELATED WORKS

Before deep learning was widely used, researchers used traditional bottom-up approaches for text detection tasks, which were broadly classified into two categories: connected components(CC) analysis based approaches [3]–[5] and sliding window based approaches [6]–[8]. CC analysis based approaches include edge detection and text-level detection. The former detects edge or corner of text instance to obtain text candidate region, which uses operators such as Canny, Sobel, or K-means clustering method. The latter detects connected region to get text candidate region, and the representative methods include stroke width transform(SWT) [17] and maximally stable extremal regions(MSER) [5]. Sliding window based approach scans the entire image by a sliding window, regards the region covered by each detection window as a text candidate region, then extracts the manully designed features within. The confidence of the candidate region is obtained by a well-trained classifier, by comparing the confidence with the threshold, the candidate region is classified into text region or background.

With the continuous development of deep learning, approaches [13]–[15], [18]–[27] based on deep learning have gradually demonstrated superiority in text detection. These methods are divided into three categories: region proposal based, segmentation based and hybrid methods. Region proposal based methods regress bounding box of text instance as the final detection result, which are divided into one-stage methods and two-stage methods. One-stage methods directly predict the bounding box of text region, mostly based on the general object detection framework SSD [28], including Textboxes [20], Textboxes++ [21], RRD [22], SegLink [23], etc. Two-stage methods first generate region proposals from feature maps, and then classify and regress these proposals to get final bounding boxes. Most of them are based on the general object detection framework Faster R-CNN [29], including RRPN [13], R2CNN [14], SLPR [15], etc. Segmentation based methods treat text detection as a generalized segmentation task, using typical semantic segmentation methods to perform pixel-level text/background annotation. PixelLink [24] adds eight-directions link prediction as the same time as the text/background classification, and generates the bounding box directly from the link. FTSN [25] uses multi-scale fused feature map and uses multi-object joint training of pixel prediction and edge detection. Inceptext [26] proposes the Inception-Text module and deformable PSROI module to detect multi-oriented text. Reference [27] exploits bootstrapping for data augmentation and semantics-aware text border segments, in order to get complete and accurate text lines. Hybrid methods use Mask R-CNN framework [16], which combines object detection and segmentation tasks and achieves better results without any other tricks.

Although more and more methods have achieved advanced performance in multi-oriented text detection tasks, they are not ideal for arbitrary shape text detection. Reference [10] proposes a curved text dataset CTW-1500, which uses

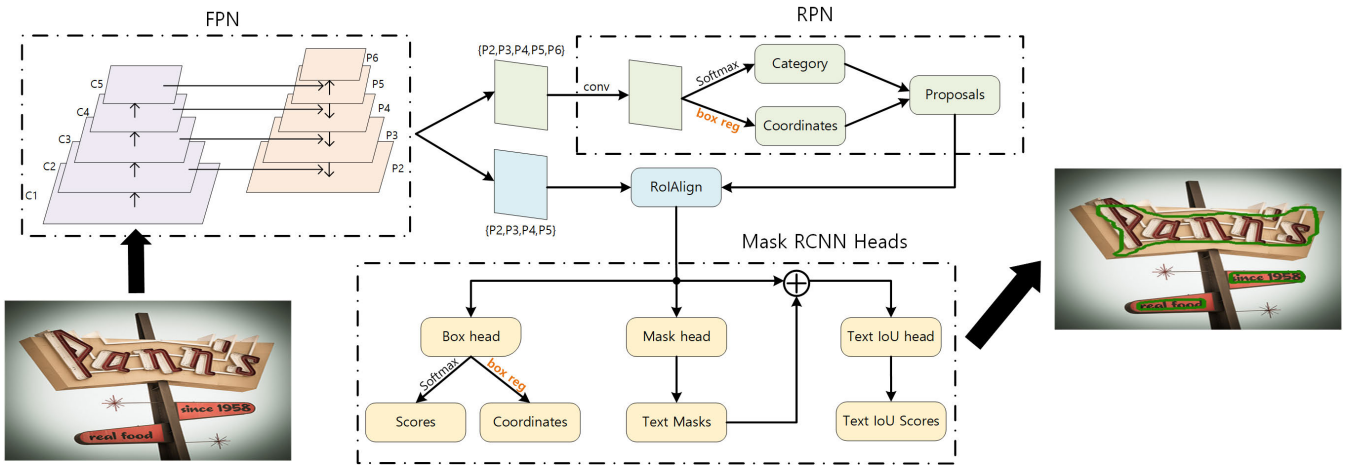


FIGURE 1. The architecture of our method. Given an image as input, the Feature Pyramid Network (FPN) extracts features from different layers and produces feature maps of different size. Region Proposal Network (RPN) uses features from P_2 to P_6, in which P_6 is the result of P_5 upsampling. While RoIAlign needs proposals from RPN and features from P_2 to P_5. Then three different heads are connected to RoIAlign. Box head predicts text/background confidence scores and coordinates of the text proposals. Mask head predicts pixel-level text mask. Text IoU head predicts IoU of text masks provided by mask head. It is worth noting that the ℓ_1-smooth loss in box reg process marked in orange is replaced by the IoU-related loss proposed in our method.

14 points to annotate the polygon ground truth, and proposes curve text detector (CTD) and recurrent trans-verse and longitudinal offset connection (TLOC) algorithm to detect curved text instances. TextSnake [30] concatenates a series of disks to describe text instances, and uses FCN+FPN network to predict text region (TR), text center line (TCL), angle and radius. In LOMO [31], direct regression (DR) module generates text proposals firstly, then iterative refinement module (IRM) refines the quadrangle proposals neatly close to ground truth by regressing the coordinate offsets once or more times. Finally, shape expression module(SEM) regresses the geometry attributes of text instances. ATRR [32] is the first to use adaptive number of pairwise points to represent text, and uses RNN+LSTM to detect arbitrary shape text. Compared with Mask R-CNN based methods, ATRR needs less computation.

In previous Mask R-CNN based methods, there are two problems that are often ignored: One is that there is not a strong corelation between minimizing ℓ_1-smooth loss and improving the IoU value between the proposed bounding box and the corresponding ground truth. The other is that there is not a strong corelation between the classification confidence score and the text mask IoU value. Text detection is a sub-area of object detection. Improving its detection performance is to improve the IoU between the detection area and the corresponding ground truth, because the determining factor during calculating the recall and precision in the evaluation stage is IoU. Therefore, our method has made corresponding adjustments based on Mask R-CNN for the above mentioned problems: For the first problem, we replace the ℓ_1-smooth loss with the proposed IoU-related loss function. During the backpropagation of the neural network, we can directly improve the IoU between predicted bounding box and the corresponding ground truth by minimizing this loss. For the second problem, we add the text scoring branch which predicts the IoU between text mask and the corresponding ground truth, and use this IoU as a determining factor of the confidence score in the final evaluation stage. Our method is simple but effective, and the performance is improved without introducing other modules to increase model computation.

III. THE PROPOSED METHOD

A. NETWORK ARCHITECTURE

The architecture of our method is shown in Fig.1. Based on Mask R-CNN benchmark, our text detection model is divided into 4 parts: the Feature Pyramid Network(FPN), the Region Proposal Network(RPN), the RoIAlign and the Mask R-CNN heads.

In convolutional neural network, deep features contain more semantic information, while shallow features contain more location information. Therefore, it is necessary to use both deep and shallow features to meet the needs of classification and localization. FPN is the backbone of the proposed model to improve the problem of multi-scale object detection. It contains three parts: a bottom-up connection, a top-down connection and a lateral connection, as shown in the dashed box labeled FPN in Fig.1. The bottom-up connection uses feature maps divided into 5 stages according to its size. Each stage takes the last residual block of Resnet-50, together form $\{C_2, C_3, C_4, C_5\}$, whose stride relative to the original image is $\{4,8,16,32\}$. In order to save memory, conv1 of stage 1 is discarded. The top-down pathway upsamples from higher-level and semantically stronger feature maps, so as to get higher resolution features. The lateral connection fuses the upsampled result with the same-sized feature map generated from the bottom-up connection. Specifically, after the conv1$\times$1 operation performed for each stage in $\{C_2, C_3, C_4, C_5\}$, they are summed with upsampled feature maps. Then conv3$\times$3 is performed to eliminate the aliasing effect of the upsampling process. As a result, we get a set of feature maps,

denote as $\{P_2, P_3, P_4, P_5\}$. P_6 is added into the feature pyramid in order to get a larger anchor scale, where P_6 is the result of stride two upsampling of P_5. In summary, Resnet-FPN uses feature maps of $\{P_2, P_3, P_4, P_5, P_6\}$ as input of RPN, while $\{P_2, P_3, P_4, P_5\}$ as input of RoIAlign.

The RPN part generates anchors of different scales on the feature maps of $\{P_2, P_3, P_4, P_5, P_6\}$. During training, after classification and bounding box regression, 2000 proposals are selected as another input of RoIAlign. This process is the first stage of the two-stage object detection model.

The RoIAlign part is an improvement on the RoI Pooling of Faster R-CNN [29]. It solves the problem that the latter feature map is not pixel-aligned with the original image and affects the detection accuracy [16]. RoIAlign is used to extract the RoI features of the proposals predicted by RPN, and normalize the size of the RoI features to the size of the Mask R-CNN heads' input, thus speeding up the training and inference process.

Mask R-CNN heads are the second stage of the two-stage object detection model. In addition to the bounding box classification and regression in Faster R-CNN, the mask head is added to predict the binary mask of text area. The text IoU head is also added to predict the IoU between text mask and corresponding ground truth as the text IoU score to affect the final confidence score. A detailed description of the text IoU head is shown in text scoring branch section.

The above four parts are combined to form a multi-task text detection model including text/background classification, bounding box regression, semantic segmentation and IoU regression.

B. GENERALIZED COMPLETENESS-AWARE IoU LOSS

The loss function used in the bounding box regression process in previous Mask R-CNN based methods is ℓ_1-smooth loss, which is calculated as follow:

$$smooth_{L_1} = \begin{cases} 0.5x^2 & if\ |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{1}$$

where x is the absolute value between the predicted bounding box regression and the ground truth regression. During back-propagation, the neural network indirectly achieves the goal of improving IoU by reducing this loss so that the predicted bounding box is closer to the ground truth. However, IoU is the metric that compares the similarity of two arbitrary two-demensional shapes. Using loss function such as ℓ_1-smooth loss does not directly improve IoU, because when two shapes overlap in different ways, ℓ_1 or ℓ_2-norms values can be the same, while the value of IoU may not be the same [33]. In order to improve the IoU directly, [33] proposed GIoU loss and used it as the loss function in replace of ℓ_1-smooth loss. Inspired by [33], we made two improvements and apply it to our network:

1) We proposed generalized completeness-aware IoU, which makes the predicted box recall the ground truth as completely as possible;

Algorithm 1 Generalized Completeness-Aware IoU Loss

Input: Predicted B^p and ground truth B^g bounding box coordinates:
$B^p = \{x_1^p, y_1^p, x_2^p, y_2^p\}$, $B^g = \{x_1^g, y_1^g, x_2^g, y_2^g\}$.

Output: $\mathcal{L}_{GCAIoU}$

1: For the predicted box B^p, ensuring $x_2^p > x_1^p$ and $y_2^p > y_1^p$:
$\hat{x}_1^p = min(x_1^p, x_2^p)$, $\hat{x}_2^p = max(x_1^p, x_2^p)$,
$\hat{y}_1^p = min(y_1^p, y_2^p)$, $\hat{y}_2^p = max(y_1^p, y_2^p)$,
2: Calculating area of B^g: $A^g = (x_2^g - x_1^g) \times (y_2^g - y_1^g)$.
3: Calculating area of B^p: $A^p = (\hat{x}_2^p - \hat{x}_1^p) \times (\hat{y}_2^p - \hat{y}_1^p)$.
4: Calculating intersection $\mathcal{I}$ between B^g and B^p:
$x_1^{\mathcal{I}} = max(\hat{x}_1^p, x_1^g)$, $x_2^{\mathcal{I}} = min(\hat{x}_2^p, x_2^g)$,
$y_1^{\mathcal{I}} = max(\hat{y}_1^p, y_1^g)$, $y_2^{\mathcal{I}} = min(\hat{y}_2^p, y_2^g)$,
$\mathcal{I} = \begin{cases} (x_2^{\mathcal{I}} - x_2^{\mathcal{I}}) \times (y_2^{\mathcal{I}} - y_1^{\mathcal{I}}) & if\ x_2^{\mathcal{I}} > x_1^{\mathcal{I}}, y_2^{\mathcal{I}} > y_1^{\mathcal{I}} \\ 0 & otherwise. \end{cases}$
5: $IoU = \frac{\mathcal{I}}{\mathcal{U}}$, where $\mathcal{U} = A^p + A^g - \mathcal{I}$.
6: $CAIoU = IoU \times ratio$, where $ratio = \frac{\mathcal{I}}{A^g}$.
7: Finding the coordinate of smallest enclosing box B^c:
$x_1^c = min(\hat{x}_1^p, x_1^g)$, $x_2^c = max(\hat{x}_2^p, x_2^g)$,
$y_1^c = min(\hat{y}_1^p, y_1^g)$, $y_2^c = max(\hat{y}_2^p, y_2^g)$,
8: Calculating area of B^c: $A^c = (x_2^c - x_1^c) \times (y_2^c - y_1^c)$.
9: $GCAIoU = CAIoU - \frac{A^c - \mathcal{U}}{A^c}$.
10: $\mathcal{L}_{GCAIoU} = 1 - GCAIoU$.

2) we applied the proposed $\mathcal{L}_{GCAIoU}$ to the bounding box regression process in both RPN and Mask R-CNN heads.

The calculation of our proposed generalized completeness-aware IoU loss is shown in Alg.1. First we calculate the area of the predicted bounding box and the ground truth box together with their IoU. Next, multiply the IoU by a ratio to get the $CAIoU$, as shown in Alg.1(6). This multiplier is the ratio of the intersection and the ground truth box area, which is used to perceive whether the predicted box can recall the complete ground truth box. As shown in Fig.2, $CAIoU$ has a greater penalty for predicted boxes that do not recall ground truth boxes completely. The introduction of $CAIoU$ makes the predicted box shift towards the direction of complete recall, which means that the cutting behavior can be suppressed. After that, we seek the smallest enclosing bounding box of the predicted box and the ground truth, calculate its area, and calculate $GCAIoU$, as shown in Alg.1(7-9). The introduction of $GCAIoU$ is to solve the problem that the value of $CAIoU$ and the gradient is zero in the case of non-overlapping. As can be seen from Fig.2, the value of $GCAIoU$ is still smaller than IoU and $GIoU$, proving that $GCAIoU$ is more universal while being capable of compleness awareness. Finally, we use the proposed $\mathcal{L}_{GCAIoU}$ in the bounding box regression process as a loss.

$GCAIoU$ is a metric closely related to IoU. Using this metric as the loss of box regression branch is an optimal choice to directly improve IoU. Applying L_{GCAIoU} to the regression branch in RPN can make the filtered proposals recall more content of the ground truth. To our knowledge,

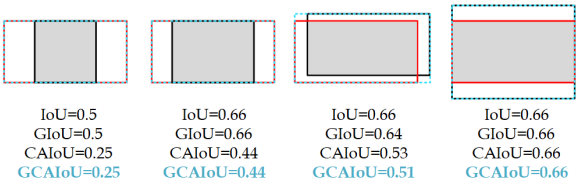


FIGURE 2. A set of examples of predicted bounding box and corresponding ground truth. For the left three cases, the predicted bounding box does not recall the ground truth completely, so as a penalty, the value of CAIoU and GCAIoU are smaller than IoU and GIoU. While for the far right case, the value of CAIoU and GCAIoU are the same as IoU and GIoU due to the complete recall. Black: detection. Red: ground truth. Grey: Intersection between them. Blue: the smallest enclosing box of them.

few works focus on the optimization of the RPN part. Our method improves the quality of the RPN proposals in a simple but effective way. Applying L_{GCAIoU} to the box regression branch of Mask R-CNN heads can further improve the performance of regression.

C. TEXT SCORING BRANCH

During the evaluation phase in current Mask R-CNN based methods, classification confidence score predicted by classification branch of box head in Mask RCNN heads plays an important role to filter redundant bounding boxes. However, since this score is not directly related to the quality of the text mask, high mask IoU but low classification confidence boxes are filtered out, thus retaining suboptimal boxes. To overcome this problem, inspired by [34], the text scoring branch is introduced to adjust the final confidence score to align it with the quality of the text mask.

The architecture of text scoring branch is shown in Fig.3. The input of Text IoU head is the concat of RoI features from RoIAlign and text masks predicted by mask head. During concatenating, a stride two maxpooling layer with a size-two kernel is used to keep the size of text mask to be the same with RoI feature. The Text IoU head includes four convolutional layers and three fully connected layers. The output of the last fully connected layer is the predicted IoU between text mask and ground truth. This IoU value is then multiplied by the confidence score predicted by classification branch to obtain the final score, denoted as $s_{\text{text}} = s_{\text{cls}} \cdot s_{\text{textIoU}}$. Thus, this score can represent the quality of the text mask for the subsequent filtering process.

D. MULTI-TASK LOSS FUNCTION

Our proposed method is a multi-task text detection model consisting of text/background classification, bounding box regression, semantic segmentation and IoU regression. We define its multi-task loss function as follows:

$$L = L_{cls} + \lambda_1 L_{box} + \lambda_2 L_{mask} + \lambda_3 L_{textIoU} \tag{2}$$

where λ_1, λ_2 and λ_3 are the weights of the bounding box regression loss, the semantic segmentation loss and the TextIoU loss, respectively. We use Softmax loss as the classification loss, which is calculated as follow:

$$L_{cls}(p_i, p_i^*) = \frac{1}{i} \sum_i -log\left[p_i^* p_i + (1 - p_i^*)(1 - p_i)\right] \tag{3}$$

where i is the index of the anchors, p_i is the predicted probability of detection, p_i^* is the probability of ground truth which is 0 or 1 otherwise.

The bounding box regression loss is our proposed L_{GCAIoU}, which is calculated in Alg.1.

Following Mask R-CNN, the loss of semantic segmentation is the average binary cross-entropy loss defined as follow:

$$L_{mask} = -\frac{1}{m^2} \sum_{1 \le i,j \le m} [\hat{y}_{ij}^* log y_{ij} + (1 - \hat{y}_{ij}^*) log(1 - y_{ij})] \tag{4}$$

where $\hat{y}_{ij}^*$ is the label of a cell (i, j) in the ground truth mask for the RoI region size of $m \times m$, y_{ij} is the predicted value of the same cell.

Following [34], we use ℓ_2 loss for regressing TextIoU, which is calculated as follow:

$$L_{textIoU}(t_i, t_i^*) = \frac{1}{i} \sum_{i \in \{y,h\}} \left| t_i - t_i^* \right|^2 \tag{5}$$

where t_i is the predicted IoU of text mask, t_i^* is the ground truth mask IoU.

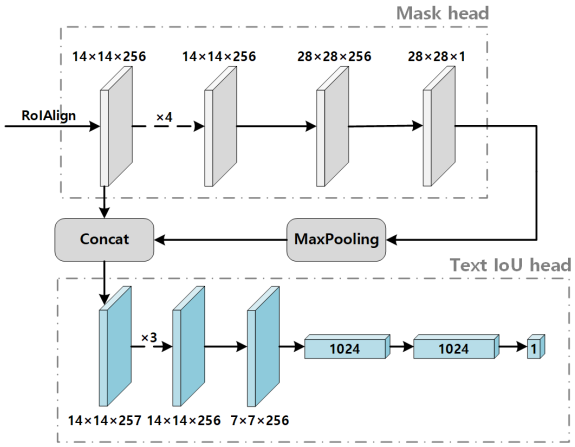


FIGURE 3. The architecture of text scoring branch. We concat the RoI feature and the predicted text mask as the input of Text IoU head. After 4 convolutional layers and 3 FC layers, the output of Text IoU head is the predicted IoU between text mask and corresponding ground truth.

IV. EXPERIMENTS AND ANALYSIS

A. DETAILS

We choose Pytorch as our deep learning framework to implement our method. We use Resnet-50 for feature extraction, synthetic data [35] for pretraining the model, and provided training data from CTW-1500 [10], Total-Text [11], ICDAR2015 [9] and ICDAR2017-RCTW [12] for fine-tuning our model.

The maximum iteration number is 60 epochs for each dataset on two Nvidia Titan X GPUs. We set the initial

learning rate to 10^{-2} and reduced to 10^{-3} and 10^{-4} on the 40th and 50th epoch, respectively. The weight of text scoring branch was set to 3.0 so as to make a balance between all branches. To prove the superiority of our method, we did not use tricks such as online hard example mining(OHEM) and data augmentation strategies. In order to save computing resources and improve efficiency, our training images were rescaled to fixed size with the width of less than 1280 pixels and the height of less than 720 pixels.

B. LABELS

In the RPN training process, whether the label of anchor is positive or negative is determined by the IoU between the anchor and the corresponding ground truth.

A positive label is defined as:

1) An anchor that has an IoU overlap higher than 0.7 with any ground truth box;
2) An anchor with the highest IoU overlap with a ground truth box.

While a negative label is defined as: An anchor has an IoU overlap less than 0.3 with all ground truth boxes.

For anchors with IoU between 0.3 and 0.7, they would be discarded rather than feeding to the heads.

C. DATASETS

Our proposed method is evaluated on three public datasets. The first two are arbitrary shape text datasets, while the latter one is a multi-oriented text dataset.

CTW-1500 [10]: It is the first curve text dataset that pioneered the field of arbitrary shape text detection. It contains 1000 scene images for training and 500 for testing, with 10,751 bounding boxes (3,530 are curve bounding boxes). Text instances are annotated by a polygon consisting of 14 points.

Total-Text [11]: It is a scene text dataset collected with multi-lingual curved text. It contains 1255 scene images for training and 300 for testing, with 9,330 annotated words. The orientations of text in this dataset is multivariate, including horizontal, multi-oriented and curved. Text instances are annotated by adaptive number of points and are labeled at word level.

ICDAR2015 [9]: It is a popular incidental scene text dataset which contains 1000 scene images for training and 500 for testing. In contrast to the two datasets mentioned above which are well-captured, images in ICDAR2015 are captured without any specific prior action, so some of them are distorted and out of focus. Text instances are annotated by the upper left and lower right corners of multi-oriented bounding boxes.

ICDAR2017-RCTW [12]: It is a large-scale scene text dataset focused on reading Chinese text in the wild which contains 8034 scene images for training and 4229 for testing. It fills the gap of Chinese scene text detection. Text instances are annotated by the upper left and lower right corners of horizontal or vertical bounding boxes.

To evaluate the results of our proposed model on multiple public datasets, we use ICDAR2015 IoU metric, which follows Pascal VOC [36]. This metric includes precision, recall, and H-mean, where H-mean is the harmonic mean of precision and recall.

D. RESULTS ON CTW-1500

Our proposed method is tested on curved text dataset CTW-1500, and the visualization of detection results is shown in Fig.4(a). As can be seen from the two sets of comparison figures on the left, due to the introduction of text scoring branch, our method can suppress false positives such as the texture in the middle of the airplane and the texture on the top of the yellow board. Zoom in to view small-scale text detection results. From the two figures in the middle, we can see that our method can distinguish adjacent text lines very well. And from the two figures on the right, our method is well adapted to text instances of different scales. The comparison of our method with others is shown in Table 1. We achieved 78.5% precision, 79.9% recall and 79.2% H-mean.

TABLE 1. Results on CTW-1500 dataset.

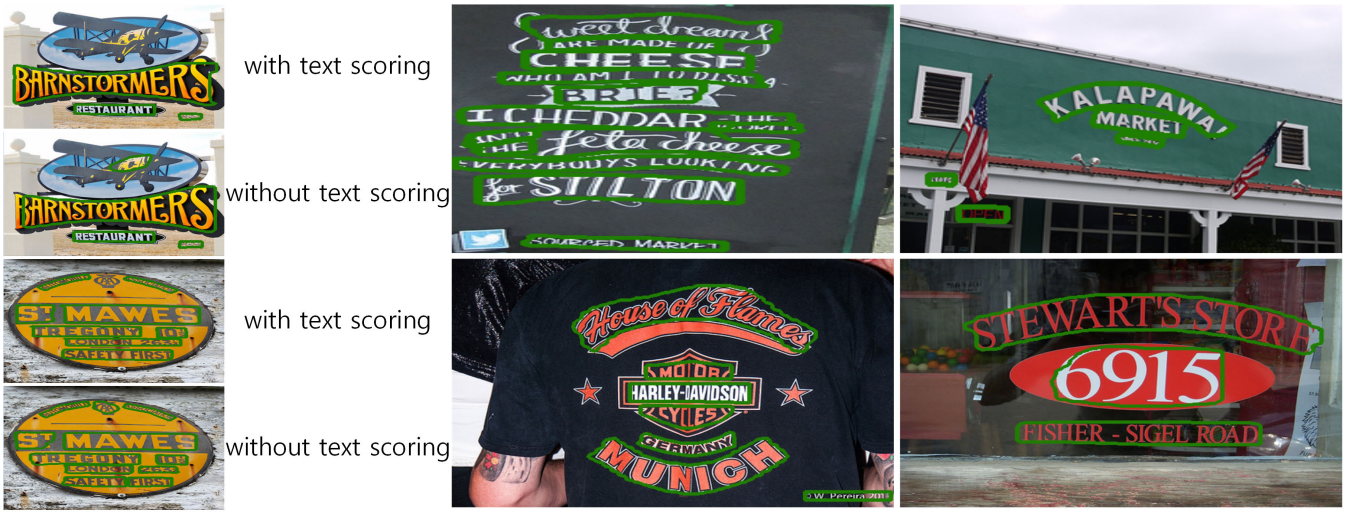
Algorithm	Precision	Recall	H-mean
ATRR [32]	80.1	80.2	**80.1**
LSE [37]	82.7	77.8	**80.1**
Proposed	78.5	79.9	79.2
LOMO [31]	69.6	**89.2**	78.4
CSE [38]	76.1	78.7	77.4
TextSnake [30]	**85.3**	67.9	75.6
SLPR [15]	70.1	80.1	74.8
CTD+TLOC [10]	69.8	74.3	73.4
CTD [10]	65.2	74.3	69.5
DMPNet [39]	56.0	69.9	62.2
EAST [40]	49.1	78.7	60.4
SegLink [23]	40.0	42.3	40.8

E. RESULTS ON TOTAL-TEXT

Total-Text is another curved text dataset with well-captured streetscapes that contains text instances of different scales, orientations and languages. The visualization of detection results on Total-Text is shown in Fig.4(b). As we can see from the figure, text instances with large bend angles and irregular shapes are well detected. We compare our detection result with other advanced methods in the past three years in Table 2. We achieved 82.1% precision, 80.0% recall and 81.1% H-mean, which is a competitive result on Total-Text dataset. Similar to the results on CTW-1500 dataset, although our method does not achieve the highest precision and recall, the two are still higher than most of other methods.

F. RESULTS ON ICDAR2015

To verify the scalability and robustness of our method, we also test on the multi-oriented scene text dataset



(a)

(b)

(c)

FIGURE 4. Visualization of detection results on different datasets. (a) Results on CTW-1500; (b) results on Total-Text; (c) results on ICDAR2015.

TABLE 2. Results on total-text dataset.

Algorithm	Precision	Recall	H-mean
ICG [41]	82.1	80.9	**81.5**
FTSN [25]	**84.7**	78.0	81.3
Proposed	82.1	80.0	81.1
PSENet-1s [42]	84.0	78.0	80.9
TextField [43]	81.2	79.9	80.6
CSE [38]	81.4	79.7	80.2
ATTR [32]	80.9	76.2	78.5
TextSnake [30]	74.5	**82.7**	78.4
TextNet [44]	68.2	59.5	63.5
Textboxes [20]	62.1	45.5	52.5
EAST [40]	36.2	50.0	42.0

TABLE 3. Results on ICDAR2015 dataset.

Algorithm	Precision	Recall	H-mean
Proposed	71.8	**89.2**	**79.6**
WordSup [45]	77.0	79.3	78.2
SSTD [46]	73.9	80.2	76.9
MCN [47]	**80.0**	72.0	76.0
SegLink [23]	76.8	73.1	75.0
FTPN [48]	68.2	78.0	72.8
DMPNet [39]	73.2	68.2	70.6

ICDAR2015. Images in this dataset are streetscapes that are not well focused. Since our method can solve the problem of detecting arbitrary shape text, it can detect multi-oriented text represented by rectangular boxes as well. The visualization of the detection results is shown in Fig.4(c). Zoom in to view small-scale text detection results. We achieved 71.8% precision, 89.2% recall and 79.6% H-mean as shown in Table 3. The recall and H-mean are higher than other representative methods among them.

G. RESULTS ON ICDAR2017-RCTW

As the scale of the previous three datasets are small, for the complete comparison of our proposed method, we also test on large-scale dataset such as ICDAR2017-RCTW. Images in this dataset focus on reading Chinese text in the wild. The visualization of the detection results is shown in Fig.5. Zoom in to get better view. We achieved 71.9% precision, 55.3% recall and 62.5% H-mean as shown in Table 4.

H. GCAIoU SUPERIORITY STUDY

The L_{GCAIoU} proposed in this paper is based on the improvement of [33]. Therefore, in order to verify the superiority of L_{GCAIoU}, we set up two sets of comparative experiments.

The first set of experiments is the H-mean comparison of three losses including the ℓ_1-smooth loss used in the Mask RCNN framework, the L_{GIoU} proposed in [33] and the L_{GCAIoU} in this paper, as shown in Table 5. It is worth noting that IC15 in the table is an abbreviation

FIGURE 5. Visualization of detection results on ICDAR2017-RCTW.

TABLE 4. Results on ICDAR2017-RCTW dataset.

Algorithm	Precision	Recall	H-mean
gmh [12]	70.6	**57.8**	**63.6**
Proposed	71.9	55.3	62.5
SARI FDU RRPN v1 [12]	71.2	55.5	62.4
LOMO [31]	**80.4**	50.8	62.3
RRD [22]	72.4	45.3	55.7
TDN SJTU2017 [12]	64.3	47.1	54.4
EAST [40]	59.7	47.8	53.1
Baseline [12]	76.0	40.4	52.8
TH-DL [12]	67.8	34.8	46.0
linkage-ER-Flow [12]	44.5	25.6	32.5

TABLE 5. H-mean comparison of three losses.

Loss	CTW-1500	Total-Text	IC15	IC17
ℓ_1-smooth loss	78.29	80.46	78.75	61.18
L_{GIoU}	78.68	80.78	78.26	61.83
L_{GCAIoU}	**78.86**	**80.90**	**79.38**	**62.48**

TABLE 6. H-mean comparison of whether L_{GCAIoU} is applied to RPN.

L_{GCAIoU}	CTW-1500	Total-Text	IC15	IC17
not in RPN	78.58	80.81	78.84	61.93
in RPN	**78.86**	**80.90**	**79.38**	**62.48**

for ICDAR2015 dataset, and IC17 for ICDAR2017-RCTW dataset. The results show that the loss function proposed in this paper gets better performance than ℓ_1-smooth loss and L_{GIoU} on both large-scale and small-scale datasets, which proves the superiority of L_{GCAIoU}.

The second set of experiments is the H-mean comparison of whether L_{GCAIoU} is applied to the RPN part. As can be seen in Table.6, applying L_{GCAIoU} to RPN has greatly improved the performance of the model.

I. ABLATION STUDY

In order to verify the effectiveness of our proposed L_{GCAIoU} and text scoring branch, we set a series of ablation

TABLE 7. H-mean on different combination of L_{GCAIoU} and text scoring branch.

L_{GCAIoU}	TS	CTW-1500	Total-Text	IC15	IC17
-	-	78.29	80.46	78.75	61.18
✓	-	78.86	80.90	79.38	62.48
-	✓	78.57	80.54	79.15	**62.92**
✓	✓	**79.18**	**81.06**	**79.56**	62.46

experiments about their different combinations, as shown in Table 4. It is worth noting that the TS in the table is an abbreviation for text scoring. As can be seen from the table, for small-scale datasets, the introduction of L_{GCAIoU} and the addition of the text scoring branch all contribute to the improvement of the model performance. Where L_{GCAIoU} improves the quality of the anchors and predicted bounding box during backpropagation, while text scoring branch improves the performance by aligning the text mask quality with the evaluation scores. The former has a superior improvement over the latter, and the combination of the two can achieve the best results. However, for large-scale dataset, although the introduction of both L_{GCAIoU} and Text Scoring branch can improve the performance of the model respectively, the best result is not obtained by the combination of the two, but by the separate Text Scoring branch.

V. CONCLUSION AND FUTURE WORKS

In this paper, we proposed an IoU-related arbitrary shape text detection model. Based on Mask R-CNN benchmark, we replaced ℓ_1-smooth loss by a generalized completeness-aware IoU as a distance which can be computed and be used as box regression loss. In addition, we added text scoring branch as a ROI head to predict text mask IoU for final evaluation stage. Our text detection model can adapt to multi-oriented and arbitrary shape text detection tasks and achieves competitive results on multiple public datasets.

In the future, the proposed method can be improved in three aspects. First, we need to figure out how to make the IoU score predicted by text scoring branch affect the quality of the mask predicted by the mask head. Second, we plan to improve the model for more challenging public datasets, such as multilingual datasets. Finally, we consider to refine the model for end-to-end detection and recognition task.

REFERENCES

[1] Y. Zhu, C. Yao, and X. Bai, "Scene text detection and recognition: Recent advances and future trends," *Frontiers Comput. Sci.*, vol. 10, no. 1, pp. 19–36, 2016.
[2] S. Karaoglu, R. Tao, T. Gevers, and A. W. M. Smeulders, "Words matter: Scene text for image classification and retrieval," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1063–1076, May 2017.
[3] P. Shivakumara, T. Q. Phan, and C. L. Tan, "A Laplacian approach to multi-oriented text detection in video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 412–419, Feb. 2011.
[4] G. Zhou, Y. Liu, Z. Tian, and Y. Su, "A new hybrid method to detect text in natural scene," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2605–2608.
[5] 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, 2010, pp. 770–783.
[6] S. M. Hanif, L. Prevost, and P. A. Negri, "A cascade detector for text detection in natural scene images," in *Proc. 19th Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.
[7] A. Mishra, K. Alahari, and C. V. Jawahar, "Top-down and bottom-up cues for scene text recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2687–2694.
[8] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1457–1464.
[9] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, and S. Lu, "ICDAR 2015 competition on robust reading," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, 2015, pp. 1156–1160.
[10] Y. Liu, L. Jin, S. Zhang, C. Luo, and S. Zhang, "Curved scene text detection via transverse and longitudinal sequence connection," *Pattern Recognit.*, vol. 90, pp. 337–345, Jun. 2019.
[11] C. K. Ch'ng and C. S. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, 2017, pp. 935–942.
[12] B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. Belongie, S. Lu, and X. Bai, "ICDAR2017 competition on reading chinese text in the wild (RCTW-17)," in *Proc. Int. Conf. Document Anal. Recognit.*, vol. 1, 2017, pp. 1429–1434.
[13] W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.
[14] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, "R2CNN: Rotational region cnn for orientation robust scene text detection," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, 2018, pp. 3610–3615.
[15] Y. Zhu and J. Du, "Sliding line point regression for shape robust scene text detection," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, 2018, pp. 3735–3740.
[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.
[17] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2963–2970.
[18] F. Zhan, C. Xue, and S. Lu, "GA-DAN: Geometry-aware domain adaptation network for scene text detection and recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 9105–9115.
[19] S. Tian, S. Lu, and C. Li, "Wetext: Scene text detection under weak supervision," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1492–1500.
[20] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: A fast text detector with a single deep neural network," in *Proc. 21st AAAI Conf. Artif. Intell.*, 2017.
[21] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3676–3690, Aug. 2018.
[22] M. Liao, Z. Zhu, B. Shi, G.-S. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5909–5918.
[23] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2550–2558.
[24] D. Deng, H. Liu, X. Li, and D. Cai, "PixelLink: Detecting scene text via instance segmentation," in *Proc. 22nd AAAI Conf. Artif. Intell.*, 2018.
[25] 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)*, 2018, pp. 3604–3609.
[26] 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," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, pp. 1071–1077.
[27] C. Xue, S. Lu, and F. Zhan, "Accurate scene text detection through border semantics awareness and bootstrapping," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 355–372.
[28] 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.* Amsterdam, The Netherlands: Springer, 2016, pp. 21–37.

[29] 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.
[30] 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.
[31] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Apr. 2019, pp. 10552–10561.
[32] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6449–6458.
[33] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 658–666.
[34] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring R-CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6409–6418.
[35] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2315–2324.
[36] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2014.
[37] Z. Tian, M. Shu, P. Lyu, R. Li, C. Zhou, X. Shen, and J. Jia, "Learning shape-aware embedding for scene text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4234–4243.
[38] Z. Liu, G. Lin, S. Yang, F. Liu, W. Lin, and W. L. Goh, "Towards robust curve text detection with conditional spatial expansion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 7269–7278.
[39] Y. Liu and L. Jin, "Deep matching prior network: Toward tighter multi-oriented text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1962–1969.
[40] 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.*, Jul. 2017, pp. 5551–5560.
[41] J. Tang, Z. Yang, Y. Wang, Q. Zheng, Y. Xu, and X. Bai, "Detecting dense and arbitrary-shaped scene text by instance-aware component grouping," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106954.
[42] X. Li, W. Wang, W. Hou, R.-Z. Liu, T. Lu, and J. Yang, "Shape robust text detection with progressive scale expansion network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 9336–9345.
[43] 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 Trans. Image Process.*, vol. 28, no. 11, pp. 5566–5579, Nov. 2019.
[44] Y. Sun, C. Zhang, Z. Huang, J. Liu, J. Han, and E. Ding, "TextNet: Irregular text reading from images with an end-to-end trainable network," in *Proc. Asian Conf. Comput. Vis.* Perth, WA, Australia: Springer, 2018, pp. 83–99.
[45] H. Hu, C. Zhang, Y. Luo, Y. Wang, J. Han, and E. Ding, "WordSup: Exploiting word annotations for character based text detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4940–4949.
[46] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3047–3055.
[47] Z. Liu, G. Lin, S. Yang, J. Feng, W. Lin, and W. L. Goh, "Learning Markov clustering networks for scene text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, May 2018, pp. 6936–6944.
[48] F. Liu, C. Chen, D. Gu, and J. Zheng, "Ftpn: Scene text detection with feature pyramid based text proposal network," *IEEE Access*, vol. 7, pp. 44219–44228, 2019.

FAGUI LIU received the M.S. degree from Beihang University, in 1991, and the Ph.D. degree from the South China University of Technology, in 2006. She is currently a Professor with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. Her research interests include service computing, the Internet of Things, cloud computing, and big data.

DIAN GU received the B.S. degree from Jinan University, Zhuhai, China, in 2018. She is currently pursuing the M.S. degree with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. Her research interests include computer vision, semantic segmentation, and object detection.

CHENG CHEN received the B.S. degree from Northwestern Polytechnical University, China, in 2017. He is currently pursuing the master's degree with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. His research interests include deep learning networks and computer vision.