Arbitrarily Shaped Scene Text Detection With a Mask Tightness Text Detector

Yuliang Liu, Lianwen Jin, *Member, IEEE*, Chuanming Fang

***Abstract*— Scene text in the environment is complicated. It can exist in arbitrary text fonts, sizes or shapes. Although scene text detection has witnessed considerable progress in recent years, the detection of text with complex shapes, especially curved text, remains challenging. Datasets with adequate samples to overcome the problem presented by curved text (or other irregularly shaped text) have been introduced only recently; however, the performance of the reported methods on these datasets is unsatisfactory. Therefore, detecting arbitrarily shaped text remains a challenging. This motivated us to propose the Mask Tightness Text Detector (Mask TTD) to improve text detection performance. Mask TTD uses a tightness prior and text frontier learning to enhance pixel-wise mask prediction. In addition, it achieves mutual promotion by integrating a branch for the polygonal boundary of each text region, which significantly improves the detection performance of arbitrarily shaped text. Experiments demonstrate that Mask TTD can achieve state-of-the-art performance on existing curved text datasets (CTW1500, Total-text, and CUTE80) and three common benchmark datasets (RCTW-17, MSRA-TD500, and ICDAR 2015). It is worth mentioning that on CTW1500, our method can outperform previous methods, especially at higher intersection over union (IoU) thresholds (16% higher than the next-best method with an IoU threshold of 0.8), which demonstrates its potential for tight text detection. Moreover, on the largest Chinese-based dataset RCTW-17, Mask TTD outperforms other methods by a large margin in terms of both the Average Precision and F-measure, showing its powerful generalization ability.**

***Index Terms*— Scene text, arbitrarily shaped text, mask, universal, convolutional neural network, CTW1500 dataset, label, text detection, segmentation.**

I. Introduction

SCENE text, which is omnipresent in everyday life, conveys valuable information. The automation of scene text detection would be useful for many promising applications, such as real-time multilingual translation, robot navigation, and image retrieval. Recognition of the information conveyed by the text requires the scene text to be detected in advance.

Unlike general objects that have a specific appearance, the various characteristics of scene text, including the length, font, size, and shape of the text, mean that scene text detection is a unique and challenging problem. Although many horizontal and multi-oriented text detection methods have achieved significant progress in recent years, nearly all state-of-the-art methods perform unsatisfactorily in detecting scene text with more complex shapes, especially curved text or text with perspective distortion. However, text incidentally captured using a smartphone usually presents with perspective distortion, and curved text is also common in the real world, such as the text on most types of columnar objects (bottles, stone piles, etc.), spherical objects, complicated planes (clothes, streamers, etc.), coins, logos, and signboards. A universal text detector should be sufficiently robust to localize arbitrarily shaped scenes. Linking methods [1]–[4] can detect components of the text and then group them together to match the text region. However, a large amount of text stacked up together makes it extremely difficult to apply heuristic empirical connection rules to group the tiny components properly; in one way or another, these methods always produce more false positives compared to direct detection methods in practice.

In general, the main challenges associated with complex-shaped text detection are as follows:

1) **Shape diversity.** Complex-shaped text is highly diverse (e.g., wavy, circular, oval, stylized, and distorted text.) A traditional rectangular or quadrilateral bounding box can only loosely localize such text, and a rigid bounding box may cause severe mutual interference.

2) **Limited data.** Most benchmark datasets contain very few samples of complex-shaped text, especially curved text. Thus far, only three datasets that contain curved text have been established: CTW1500 [5], Total-text [6], and CUTE80 [7]. The first two were only released recently, and the performance of the reported methods on these datasets is unsatisfactory. Details of these datasets reported in Section IV.

In this paper, to detect arbitrarily shaped scene text, we propose a conceptually simple method named Mask TTD. To enhance text mask prediction, we propose a tightness prior method for processing region proposals; thus, the entire text region can be conserved during training. Furthermore, a frontier text mask is learned and integrated into mask prediction, which can split stacked text effectively. Inspired by previous methods [5], [8], our method further utilizes polygonal boundary regression to improve the prediction performance of the mask. The branches operate on the same Regions of Interest (RoIs) to perform instance segmentation and share the same backbone of the convolutional neural network (CNN). We find that such branched mutual promotion combines the advantages

Manuscript received March 20, 2018; revised January 11, 2019 and September 11, 2019; accepted November 2, 2019. Date of publication November 26, 2019; date of current version January 28, 2020. This work was supported in part by the NSFC under Grant 61936003 and Grant 61673182, in part by the National Key R&D Program of China under Grant 2016YFB1001405, in part by GD-NSF under Grant 2017A030312006, and in part by Guangzhou Science and Technology Plan (GZSTP) under Grant 201704020134. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sen-Ching Samson Cheung. *(Corresponding author: Lianwen Jin.)*

The authors are with the School of Electronics and Information Engineering, South China University of Technology, Guangzhou 510640, China (e-mail: lianwen.jin@gmail.com).

Digital Object Identifier 10.1109/TIP.2019.2954218

of direct regression and Fully Convolutional Network (FCN)-based algorithms [9], an approach that is experimentally shown to improve text detection performance.

Experiments on all existing curved text datasets, including CTW1500 [5], Total-text [6], and CUTE80 [7] showed that Mask TTD significantly outperforms previously reported methods, especially in the case of higher Intersection Over Union (IoU) thresholds (0.7, 0.8) on the CTW1500 dataset, which demonstrates the robustness of Mask TTD for tightly detecting curved text. Moreover, Mask TTD also achieves state-of-the-art performance on well-known datasets, e.g., the ICDAR 2015 challenge 4 [10], MSRA-TD500 [11], and RCTW-17 [12]. Especially on RCTW-17, our method can outperform the previous methods by a large margin on both F-measure and Average Precision, showing its powerful generalization ability.

We summarize our contributions as follows:

- We propose a systematic framework, namely Mask TTD, which can tightly localize arbitrarily shaped text.
- We propose a novel tightness prior instance segmentation method, which dynamically adjusts text proposals to cover the entire text region and skillfully utilizes text frontier information to enhance text mask prediction.
- We propose a novel mutual branch promotion method to improve the text detection performance, which combines the advantages of both direct regression and FCN-based methods.
- We propose a simple but effective polygonal generation algorithm to transfer the mask score map into a polygon (including a rectangle and quadrangle), which is a prerequisite to evaluate the detection performance on all text datasets.
- The proposed Mask TTD can achieve state-of-the-art performance on both curved and non-curved datasets.

II. Related Work

Recently, the emergence of many scene text datasets, which are constructed for specific tasks and scenes, contributed significantly to the advancement of text reading methods [44]. One of the main reasons for such progress is the evolution of the benchmark dataset: as the data become more complex, the numbers become larger and the labels become tighter. Since 2003, rectangular labeled datasets, such as ICDAR'03 [13], ICDAR'11 [45], ICDAR'13 [46], as well as COCO-Text [47], have attracted considerable attention in studies on detection task. Since 2010, multi-oriented datasets with rotated rectangular labels, such as NEOCR [48], OSTD [20], MSRA-TD500 [11] and USTB-SV1K [4]), have emerged. These datasets have stimulated the development of many influential multi-oriented detecting methods. In 2015, the first quadrilateral labeled dataset, namely Incidental Scene Text [10], was presented at ICDAR 2015. It has attracted considerable attention according to its website [10], and it has been the focus of many recent studies. Subsequently, larger and more challenging quadrilateral labeled datasets were presented at ICDAR 2017, such as RCTW-17 [12] (large dataset for Chinese and English text), DOST [49] (scene texts observed by video in the real environment) and MLT [50] (large dataset for multi-lingual text). These datasets have now become mainstream datasets. The first dataset containing curved text, namely CUTE80, was introduced in 2014. However, it was the emergence of two subsequent curved text datasets, namely CTW1500 [5] and Total-text [6], late in 2017 that highlighted the importance of curved text detection, which has facilitated several polygonal-based methods.

The development of detection methods show a similar evolution tendency to that of methods designed for dataset annotation (from a horizontal rectangle to a rotated rectangle, quadrangle and polygon), as seen in Table I. Since 2011, methods involving rotated rectangular bounding boxes have been proposed nearly every year. 2017 witnessed the emergence of various quadrilateral based detection methods [2], [3], [38]–[41] that can achieve the best performance for rotated datasets or horizontal datasets (by evaluating the circumscribed rectangle). Moreover, they can outperform horizontal methods for multi-oriented datasets, especially in terms of the recall rate. This is mainly because the stronger supervision in quadrilateral-based methods reduces background noise, unreasonable suppression and information loss significantly [38]. The beneficial effects of stronger supervision on detection are also observed in the case of Mask R-CNN [51], which improves the detecting results through joint training with a branch of segmentation. In addition, Li *et al.* [37] also proved that training with recognition is conducive to text detection.

III. Mask TTD

In this section, we describe the key components of Mask TTD, which is a straightforward detection framework that can capture arbitrarily shaped text in natural or digital-born images. The overall architecture of the Mask TTD is shown in Figure 1. The network is trained in two stages. The Region Proposal Network (RPN) initialized from ImageNet model is trained in advance. The trained RPN is then used to generate region proposals that are processed by the proposed tightness prior and RoIAlign for subsequent R-CNN training. With regard to the R-CNN stage, inspired by Mask R-CNN [51], we integrate the direct regression curved text detector (CTD) method proposed in [5] with mask prediction, which shares several fully connected layers with classification and circumscribed rectangle regression branches. The mask prediction is enhanced by learning the boundary of the text region, which is crucial for separating stacked text. All the branches are co-trained with mask prediction simultaneously and connected to the RoIAlign layer, which uses bilinear interpolation [52] to compute the exact values of the positions of each RoI instead of using quantization. We find such branch mutual promotion can significantly improve the performance, which will be evaluated in Section IV.

A. Network Architecture

The proposed Mask TTD utilizes end-to-end instance segmentation for text detection. This is inspired by Mask R-CNN [51], the latest object detection/segmentation framework extended from Faster RCNN, which achieves superior performance. Some components of Mask R-CNN are

TABLE I

EVOLUTION OF SCENE TEXT DETECTION METHODS. *RULE: RELY ON MANUALLY DEFINED RULES TO GROUP TEXT REGIONS. BBOX: BOUNDING BOX. WE CATEGORIZED RULE-BASED METHODS ACCORDING TO THEIR EVALUATING DATASETS. THESE METHOD MAY BE ADAPTED TO CURVED TEXT DETECTION WITH FURTHER POST-PROCESSING

Year \ Text BBox	Horizontal Rectangle	Rotated Rectangle	Quadrangle	Polygon
2003	(Lucas et al.) [13]	-	-	-
2004	(Chen and Yuille) [14]	-	-	-
2005	(Lyu, Song, and Cai) [15]	-	-	-
2006	(Liu, Goto, and) [16]	-	-	-
2010	(Wang and Belongie) [17] (Epshtein et al.) [18] (Neumann and Matas) [19]	-	-	-
2011	-	(Yi and Tian) [20] (Shivakumara et al.) [21]	-	-
2012	-	(Yao et al.) [11]	-	-
2013	(Huang et al.) [22]	-	-	-
2014	(Huang, Qiao, and Tang) [23]	(Kang et al.)*rule [24] (Yin et al.)*rule [25]	-	-
2015	(Tian et al.)*rule [26] (He et al.) [27] (Liang et al.in) [28]	(Yin et al.)*rule [4]	-	-
2016	(Tian et al.) [1] (Zhong et al.) [29] (Cho et al.)*rule [30]	(Zhang et al.) [31]	-	-
2017	(Liao et al.) [32] (Zhang et al.) [33] (Zhong et al.) [34]	(Ma et al.) [35] (Jiang et al.) [36] (Li et al.) [37]	(Liu and Jin) [38] (Shi et al.) [2] (Zhou et al.) [39] (He et al.) [40] (Dai et al.) [41] (Hu et al.)*rule [3] (Wu et al.) [8] (Deng et al.) [42]	(Ch'ng and Chen) [6] (Liu et al.) [5]
2018	-	-	-	(Zhu and Du) [43]

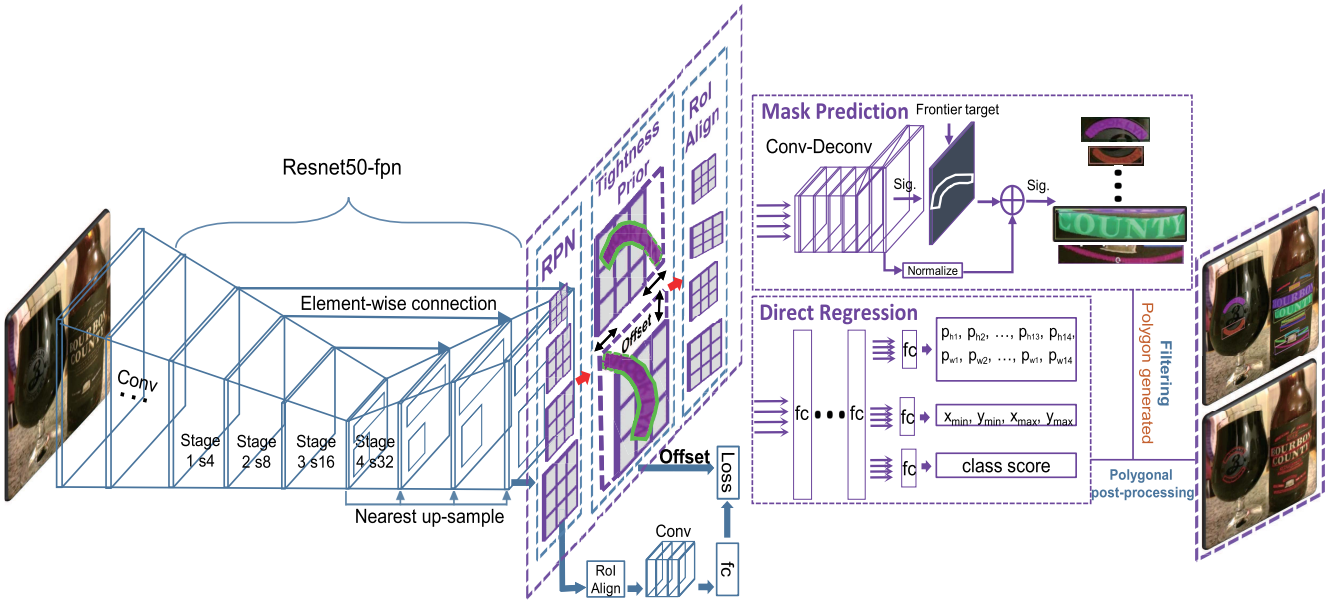


Fig. 1. Overall schematic of the Mask TTD.

utilized in Mask TTD, as they well suited for text detection: **1). RoIAlign.** RoIAlign removes quantization for aligning the extracted RoI of the feature map, which is much more accurate for predicting pixel-accurate masks than the original RoIPool [53]. This feature is crucial for text detection, where a few missing pixels may lead to erroneous recognition. **2). Instance segmentation.** Instance segmentation can mitigate the foreground adhesion problem. This feature is also

important because all common recognition methods take text-line or word-level instances as input, whereas text often stacks up together, which is detrimental to recognition performance. **3). Class-specific mask.** As text is the only foreground class, per-pixel sigmoid and binary loss for generating a mask for every object without competition among classes is beneficial for distinguishing text and non-text regions.

However, because of the characteristics of scene text, Mask R-CNN have limited ability to detect text:

- The mask prediction is restricted to the inside of the proposal, and those text regions outside the proposal are regarded as background.
- Unlike object segmentation that has pixel-level annotation, text detection datasets only provide bounding box ground truths. Hence, some exception-backgrounds inside the bounding box are regarded as text to train the mask text score map, which would degrade the classification performance to some extent.
- Text benchmark datasets all require bounding box to evaluate the detecting performance, such as rectangle, quadrangle, and polygon. However, Mask R-CNN does not have an appropriate method to group the predicted mask into a bounding box. In addition, the occurrence of noise points is inevitable in mask prediction.

Compared to Mask R-CNN, Mask TTD has many key components that are sufficiently robust to address these limitations:

- Mask TTD adopts a tightness prior layer to ensure the proposal has a sufficient size of the proposal in the training phase, thus mask prediction is trained on the basis of the entire text region (detailed in Section III-B).
- Mask TTD learns the frontier of the text region in advance to enhance the subsequent text mask prediction, which not only prevents exception-backgrounds from unraveling the text region, but also improves the prediction accuracy in practice (detailed in Section III-B).
- Mask TTD utilizes simple but effective polygonal grouping and filtering methods, which can successfully transfer the mask into a polygon for subsequent evaluation (detailed in Section III-D).
- A novel mutual promotion of Mask TTD combines the advantages of both direct regression methods and FCN-based methods, which can significantly improve the text detection performance (detailed in Section III-C).

We adopt ResNet-50-fpn as our backbone, and we use the ResNet-50 pre-trained ImageNet model [54] to finetune our task. ResNet-50 [55] is a light backbone that achieves a good trade-off between speed and accuracy, and FPN [56] can efficiently capture the context information via element-wise connection, which is especially useful for recalling small text in practice. Here, we follow the same approach as that in [56], i.e., we use the nearest neighbor method, to upsample the intermediate feature maps.

The detector is separately trained in two stages, namely the RPN stage and the Refinement R-CNN phase. In the RPN stage, we use rectangular anchors of five sizes, i.e., five aspect ratios of the convolutional feature maps (1/4, 1/8, 1/16, 1/32, and 1/64 of the original input size) to roughly cover the text, and we set a loose RPN-NMS threshold to avoid the risk of premature suppression. To connect the R-CNN, we adopted RoIAlign [51] as a sampling method. RoIAlign is a quantization-free layer which can preserve the accurate spatial location, which significantly improves the mask prediction accuracy. We adopt 7×7 RoIAlign to generate feature maps from each scale of the intermediate layers for the following R-CNN. The R-CNN stage consists of four parts: mask prediction, classification, circumscribed boundary rectangle localization, and transverse and longitudinal polygonal offset prediction.

B. Tightness Prior

In instance segmentation, mask prediction is relatively accurate, but it is constrained by the bounding box. If the text region is not completely covered by the bounding box, the text outside the text region is ignored, which would visibly degrade both the text detection and recognition performance.

In fact, this phenomenon also occurs during the training procedure, because the region proposals generated by the RPN in the first stage may also truncate the text region. Forcing prediction of the mask of inadequate text regions would introduce more false positives (box-in-box) in practice.

To address this problem, we propose a tightness prior layer to compensate for the missing text region. During training, each proposal is assigned to one ground truth (GT) text region, enabling us to judge whether the proposal entirely covers the text. If not, the proposal is enlarged by an enhanced regression process after the original RPN regression, and the new proposal is used for subsequent R-CNN training. Contrary to the original RPN regression, the tightness prior (1) only enlarges proposals that do not cover the text entirely, rather than using a shrink operation, which is not adopted in this process. Therefore, the predicted offsets can only be positive; (2) it calculates only the offsets only inside the feature of the predicted anchor region. As shown in Figure 1, the anchor is first processed with RoIAlign followed by stacked convolutions and a fully connected layer. The approach of forcing the offsets to be learned from the inside to outside differs from that of the original RPN regression. Overall, there are four offsets for the top, right, bottom and left sides, which serve as ground truths (p^*). Further, each proposal crops a region from the source feature map and uses RoIAlign followed by convolution and fully connected layers to learn four targets (p_i), where $i \in S = \{1, 2, 3, 4\}$ which corresponds to the four sides). The loss is computed by Smooth-L1:

$$Lreg(p; p^*) = \sum_{i \in S} smooth_{L_1}(p_i, p^*), \tag{1}$$

where

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

Because the offsets can only be positive, proposals that only loosely cover the entire text are not affected by the tightness prior.

Moreover, we introduce the frontier target to enforce the text region. The frontier mask is a binary map generated

Fig. 2. Left: detection results without using Tightness Prior. Right: detection results with tightness prior. The floating-point value in the images on the right are the predicted confidences of each bounding box.

by GT following an existing strategy [8]. During training, the frontier target is learned in advance, and the predicted frontier feature map is then connected element-wise (addition) with the normalized input layer, as shown in Figure 1. Finally, the entire text region is predicted on the basis of this frontier-enhanced feature map. This procedure effectively deconstructs the stacked text region. Following [51], Mask TTD also uses a per-pixel sigmoid to activate mask prediction and applies average binary cross-entropy loss.

The function of the frontier mask can be identified as changing the learning weight of each pixel. Benefiting from the tightness prior, which guarantees the completeness of the entire text, the frontier of the text region in the training image can also be conserved completely.

Figure 2 shows some examples that highlight the difference in performance with and without tightness prior. The proposed tightness prior has good potential to tightly detect the text, which is especially useful when evaluating detection performance under high IoU conditions.

C. Branched Mutual Promotion

As shown in Figure 1, the classification and boundary rectangular branches are the same as those of Fast R-CNN [53], which uses softmax loss and smooth-L1 loss, respectively to learn the targeting value. The use of the polygonal boundary direct regression branch was inspired by a recent study [5], which proposed a polygon-based curved text detector (CTD) to separately predict width/height offsets for curved text detection. In addition, the network architecture can be seamlessly integrated with an RNN to learn the inherent connection between locating points to improve the accuracy and smoothness of the detection. CTD+TLOC [5] shows excellent performance for both curved and non-curved text, indicating that such strong supervised direct regression is conducive to text detection. Similarly, we regress the relative length w_i and h_i $(i \in 1, 2, \ldots, 14)$ of every point. However, we simultaneously predict the offsets w and h in a single branch without RNN, because all the evaluation results are based on the mask prediction branch and subsequent grouping method, whereas RNN is mainly used for smoothing the detection result. Following this approach [5] and its number of annotations for the curved text, the total number of polygonal regressing items is 28, i.e., the offset of the 14 points. The parameterizations of the offsets (d_{w_i} and d_{h_i}) are listed below:

$$\begin{cases} d_{w_i} = \dfrac{p^*_{w_i} - p_{w_i}}{w_{chr}}, \\ d_{h_i} = \dfrac{p^*_{h_i} - p_{h_i}}{h_{chr}}, \end{cases} \quad (i \in (1, 2, \ldots, 14)) \tag{3}$$

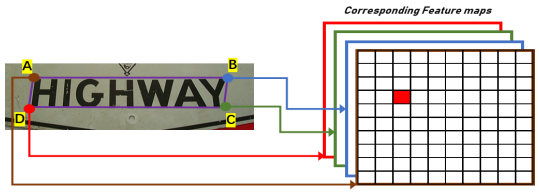
where, p^* and p are the ground truth and predicted offsets, respectively. Further, w_{chr} and h_{chr} are the width and height of the circumscribed rectangle, respectively.

The Mask branch encodes each RoI feature by conducting five convolutions with identical sizes and a deconvolution operation with stride 2, and thus the final size of the feature map is 28×28. This branch predicts a score map by pixel-wise sigmoid activation guided from aggregate frontier mask, and the prepared text-region-level mask ground truth is the binary images based on the polygonal annotating region. The crucial improvement of the Mask TTD is based on the mutual promotion between the mask prediction and direct polygonal boundary regression.

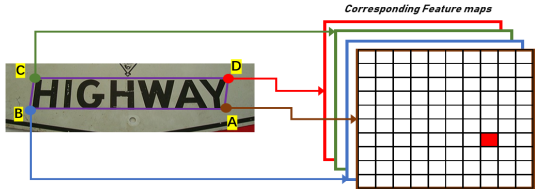
Compared to direct regression, mask prediction has the following advantages:

- It is not constrained by text shape; thus, it can effectively detect strongly curved text. By contrast, direct regression methods rely on a fixed number of points of the annotated bounding box, which is insufficient for strongly curved text.
- The text score map is trained on the basis of pixel-wise strong supervision, of which the detection results are significantly tighter than those of direct regression methods with a higher IoU threshold, as illustrated in Section IV.
- In practice, it has a higher recall rate than direct regression methods.

Moreover, to successfully train Mask TTD, the use of an appropriate label sequence is a prerequisite for mutual promotion, otherwise the results would be significantly worse. In order to to find a more effective solution to this problem, we begin by discussing the cause of the labeling sequence. This sequence or the position of points of the ground truth boxes is not relevant to mask prediction, whereas direct regression is highly sensitive to the annotation. It has been shown in [38], [43] that the sequential protocol is importance for direct regression methods, especially for some one-stage regression methods, such as [39] and [40]. For example, if pseudo samples are used by rotating the training images,



(a) Annotation 1.



(b) Annotation 2.

Fig. 3. Both (a) and (b) have exactly the same input images and networks but differ w.r.t. their annotating sequence. For (a), the first corresponding feature map is to learn the point A while for (b), because point A is at the bottom right, the same feature map learns a different target with the same input image. Therefore, confusion occurs, and to minimize the loss of both cases, as shown in Equation 4, the equality holds when $T = \frac{A_1+A_2}{2}$, where, T represents the prediction and A represents the target. All the four corresponding feature maps cause the same problems, which result in the cases shown in Figure 4.

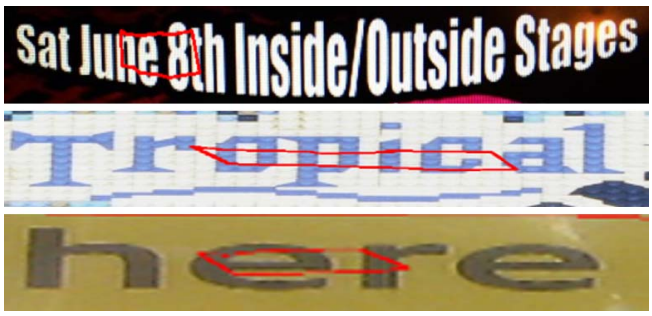


Fig. 4. Examples of concentrated bounding boxes without proper sequential pre-processing of the annotation points.

Fig. 5. Examples of labels with text reading direction. For each bounding box, the first point is marked as circle, and the remaining points follow the square direction.

the performances of [39] would be degraded significantly. Empirically speaking, this is mainly because deep-learning-based text detection relies on a broad receptive field to determine the position of the text region instead of the details. Therefore, the rotated images would easily cause the confusion of the first point and hence the remaining points.

Basically, if the first point of a bounding box can be determined, the sequence of the other points can be easily determined clockwise (or anticlockwise). However, it is difficult to establish a stable protocol to choose the first point; for example, if we decide that the point with minimum x is the first point, regarding a rectangle, the sequence 1, 2, 3 and 4 points would be top left, top right, bottom right, and bottom left (if points 1 and 4 have the same value of x, a point with a smaller value of y would be chosen as the first point). If there was another similar sample with the label of the fourth point from the left being a pixel (it is possible because annotation is subjective), the original fourth point would become the first point and the entire sequence would change because of this one unremarkable pixel (we refer to the x-min line as a choppy boundary). In such a case, the network would learn concentrated bounding boxes, as shown in Figure 4. This occurs when the network attempts to find a balanced position to minimize the ambiguous losses. For clarification, we illustrate a confusion case and provide a detailed explanation in Figure 3.

$$\begin{cases} loss_1 = |T - A_1|, \\ loss_2 = |T - A_2|, \\ loss_1 + loss_2 \geq 2\sqrt{|T - A_1| \cdot |T - A_2|}. \end{cases} \tag{4}$$

Such learning confusion would degrade the performance drastically, as is shown in Section IV. These phenomena are more likely to occur in datasets with polygonal labels, such as CTW1500 [5] and Total-text [6]. However, the annotation of many existing datasets follows the text reading direction, which is somehow easier for training direct regression methods. Text in natural images may appear in mirrored, symmetrical or retroflexed. Simply clipping the detection result of such exceptional text makes recognition difficult. If the training samples are labeled with the reading direction, as shown in Figure 5, the network would be forced to learn the reading sequence, which can be used to align the text suitably for better text recognition.

Although direct regression is disadvantage in certain respects, it also has a few remarkable advantages compared to mask prediction:

- It outputs the final localizing results without any post grouping operations (grouping methods are easily affected by noise and are less efficient), which are required to evaluate the performance of almost all existing text detection datasets.
- Direct regression methods learn the text line/word region effectively, which is less affected by adhesive stacked text than FCN-based methods.
- Direct regression methods have fewer false positives than FCN-based methods.

To exploit the advantages of both of these methods, we use multi-task loss $L = L_{cls} + L_{rectBox} + L_{mask} + L_{polyBoundary}$. Each loss is responsible for guiding each branch for the purpose of fitting the ground truth. We assume that the weakness of each branch would be complemented by another branch. For example, the outliers of mask prediction would be constrained or suppressed by the polygonal boundary regression branch. This concept is inspired by [8], which proposed a border learning method that aims to distinguish stacked text, and it significantly improves the text-block FCN performance

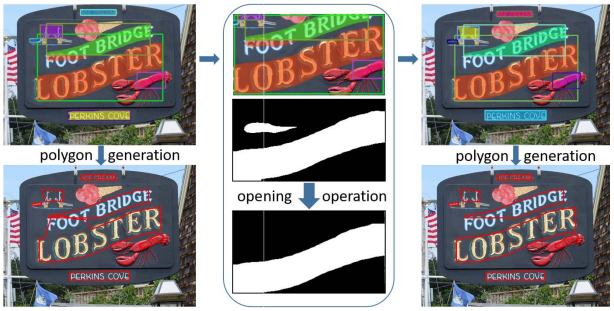


Fig. 6. Using opening operation to avoid noise during polygon generating. Left: polygon generation without filtering by opening operation. Middle: example of using opening operation. Right: polygon generation after using opening operation.

both in terms of recall rate and precision. Unlike [8] that constructed a new dataset and requires a pre-process of the training data, our method is more straightforward in learning the border through polygonal boundary regression. Moreover, instead of using FCN for segmentation of the entire image, our method is based on instance segmentation, with whole branches co-working on the same proposal.

D. Polygonal Generation With Mask Grouping

As mentioned above, Mask TTD predicts a mask for each RoI on the basis of the per-pixel sigmoid. For detection evaluation and subsequent recognition, it is necessary to group the masks into a polygon (a quadrangle or rectangle can be regarded as a special case of a polygon). Because Mask TTD is based on instance segmentation, each RoI should only contain one text region. Therefore, we utilize a sliding vertical line to generate the polygon. First, we change the mask to a (0, 1) binary map using a threshold 0.5. Then, we densely slide a vertical line to find the uppermost and lowest boundaries. Next, we evenly sample the points in the upper and lower boundaries such that a closed polygon is generated. For such grouping, the target should not have an inner bore; most examples of real scene text meet this prerequisite. However, such a grouping method is not robust against noise, which may be caused by nearby text or text-like objects. This phenomenon can be avoided by a simple but effective morphological opening operation. Empirically, we use an ellipse kernel with size (max(1, $w_{roi}/14$), max(1, $h_{roi}/14$)), and double erode operations followed by double dilate operations. Note that all positive RoIs are processed equally and this simple step can improve the accuracy without influencing the recall rate, as demonstrated in our experimental results. An example is visualized in Figure 6. The overall method to generate the polygon is described in Algorithm 1.

In addition, existing datasets, such as ICDAR 2015 challenge 4 [10] and RCTW-17 [12], require a four-point bounding box to evaluate the results. Although it is easy to find the minimum bounding area rotating the rectangle that surrounds the mask region, the rectangle is imperfect for localizing certain types of text with perspective distortion. Inspired by [57], we adopt a dynamic programming approach to find an approximate quadrilateral with minimum area that runs in

Algorithm 1 Algorithm for Polygonal Generation

Input:
M - Mask prediction feature map
L - Sliding vertical line
C - Number of the columns of M
R - Number of the rows of M
O - Set of the vertexes of the outputted polygon
i - Loop stage
γ_{ij} = the j-th pixel of the i-th L
Initialization:
(a) $O = \varnothing$
(b) C = width of the M
(c) i = 1
(d) $M = \{0,1\}$ binary map using a threshold 0.5
For $i \leqslant C$ do:
(a) Use morphological opening operation to filter M
(b) for j = 1 to R do
    if γ_{ij} == 1 then
      Assign γ_{ij} to O_i
    else
      Continue
    end if
    O_i preserves two points with MAX-y and MIN-y
    if MAX-y == MIN-y then
      MAX-y = MAX-y+1
    end if
  end for
$i = i + [C/20]$
end for
(c) Ordering points of O clockwise.
(d) (Optional) Running dynamic programing approach to generate quadrangle based on the O.

time $O(4n^3)$, where n is the number of convex hull points. To reduce the time complexity, the convex hull is found by grouping the polygon by the above-mentioned method; some examples are shown in Figure 11. As shown by the results in Table VII (last four rows), a quadrilateral bounding box can significantly outperforms a rectangular bounding box under stricter conditions, such as 0.7 or 0.8 IoU thresholds. Note that although many off-the-shelf functions exist that can generate the polygonal vertex, these methods are slower than the proposed method, as shown by the comparison in Section IV.

IV. Experiments

We tested Mask TTD on all the existing curved datasets mentioned in Section II. Then, we evaluated our method on three representative benchmark datasets to further investigate its universal applicability. All experiments were conducted using the Resnet-50-fpn backbone on MXNet [58]. The experimental environment comprised a server with Intel i7 6700K CPU, 64 GB RAM, P100 and Ubuntu 16.04 OS. PNMS [5] is used for all subsequent post-processing.

We used an alternative strategy to train the network. 1) First, the RPN was trained for 8 epochs with a model pretrained on ImageNet, using a learning rate of 0.004, which decays

Fig. 7. Example results on three curved text datasets. The last column listed some failure cases.

to 0.0004 in the sixth epoch. 2) The RPN then generated proposals to train the RCNN with ImageNet initialization. The RCNN was trained for 24 epochs using the same learning rate strategy as RPN, but the decay step was the 20th epoch. 3) Then, the RPN was trained with RCNN initialization, and we repeated the first and second steps. (4) The final model was generated by combining the well trained RPN and RCNN.

A. Experiments on All Existing Curve Text Datasets

In recent years, scene text reading has achieved significant progress, mainly owing to the powerful learning ability of deep learning algorithms. However, state-of-the-art methods are usually based on supervised deep learning, which requires a large amount of training data. Attempts to address the problems associated with reading curved text reading issue, would require a dataset containing curved text. Currently, only three such datasets exists:

- **CUTE80 [7]**. CUTE80, which is the initial curved text based dataset that was constructed in 2014, contains 80 images captured with a digital camera or retrieved from the Internet. Most images in this dataset include a small amount of text the is clean and well focused. For each line of text, the ground truth is manually annotated and contains the set of points of the polygons that forms the bounding box, and some are based on stroke level annotation. Much of the recognizable text of this dataset is unlabeled: e.g., in each of the last two images of the last column of Figure 7, only the largest text is enclosed annotated box to indicate the ground truth. This box can be used for both detection and recognition evaluation. The dataset only targets the problem of reading English text.
- **Total-text [6]**. Total-text was constructed in 2017. Compared to CUTE80, the curved text was collected from various scenes, including those with text-like scene complexity and low-contrast background. In addition, most of the images contain a large amount of non-curved text along with at least one instance of curved text, which more closely resembles real-world scenarios. The dataset includes 1,555 images (1,255 for training, and 300 for testing). The text is annotated with word-level granularity using polygons. The annotating points of the polygons are subjectively determined. Total-text is also mainly constructed to solve English reading problems.
- **CTW1500 [5]**. CTW1500 was also constructed in 2017. The dataset includes 1,500 images (1,000 for training and the remainder for testing), with text-line level annotation. Compared to the two above-mentioned datasets, its annotating method is based on relatively objectively segmented equidistant points (the point positions are partially restricted by the equal distances between the vertexes of the bounding boxes), and the fixed number of points of each curved bounding box improves its visual regularity. It can also be used for both detection and recognition evaluation. The dataset targets both English and Chinese reading problems.

TABLE II

EXPERIMENTS ON CTW1500, TOTAL-TEXT, AND CUTE80. R: RECALL RATE. P: PRECISION. H: HARMONIC MEAN OF RECALL AND PRECISION

Dataset	Algorithm	R (%)	P (%)	H (%)
CTW1500	CTD [5]	65.2	74.3	69.5
	CTD+TLOC [5]	69.8	77.4	73.4
	SLPR [43]	70.1	80.1	74.8
	Mask TTD	79.0	79.7	79.4
Total-text	Ch'ng [6]	33.0	40.0	36.0
	Mask TTD	74.5	79.1	76.7
CUTE80	Risnumawan [7]	68.0	65.0	61.0
	Mask TTD	76.9	74.3	75.6

We followed an existing protocol [5] to compute the results, i.e., we calculated the exact polygonal IoU between the ground truth and the predicted polygonal results. For fair comparison, experiments on CTW1500 and Total-text were conducted with their training sets provided without Synth-Text [59] pretraining or any data augmentation. CUTE80 only contains 80 images, which is insufficient to train our model. To overcome the insufficiency of training data, we directly used the model trained by CTW1500 to test on this dataset.

The results of the CTW1500 test set, which appear in Table II show that the Mask TTD outperformed recent state-of-the-art methods on CTW1500. We also conducted ablative studies to evaluate the effectiveness of the modules: (1) branch mutual promotion. (2) tightness prior; 3) frontier mask; (4) polygonal grouping; and (5) the influence of the label sequence. We first conducted an experiment to evaluate mutual promotion to show how mask prediction would be improved if a polygonal prediction branch is added. To understand the effect of mutual promotion on text detection, we evaluate models trained with and without the polygonal prediction branch. Specifically, in addition to the proposed models shown in Figure 1, we train additional models that adopt strictly the same architecture as Mask TTD, except for the removal of the output layer from the polygonal boundary prediction of the text region, that is, the baseline pure Mask R-CNN from TuSimple [60] repository. For a fair comparison, both evalu-

TABLE III

ABLATIVE STUDY EXPERIMENTS ON CTW1500. MT: USING BRANCH MUTUAL PROMOTION. TP: TIGHTNESS PRIOR. FM: FRONTIER MASK. GP: USING OUR GROUPING METHOD INSTEAD OF OFF-THE-SHELF OPENCV POLYGONAL GROUPING FUNCTION. OP DENOTES OPENING OPERATION. RESIZE: RESIZING THE TEST IMAGE TO 768X1024. WSQ IMPLIES THAT WE DISORGANIZE THE POINT SEQUENCE OF SOME GROUND TRUTH BOUNDING BOXES, AND STRONG AND SLIGHT IMPLY THE DISRUPTION DEGREE

	MT	TP	FM	GP	OP	Resize	WSQ (slight)	WSQ (strong)	Recall (%)	Precision (%)	Hmean (%)
baseline									75.9	66.4	70.8
baseline+	✓								76.4	77.5	76.9
baseline+	✓	✓							78.9	75.7	77.3
baseline+	✓	✓	✓						81.2	74.6	77.8
baseline+	✓	✓	✓	✓					81.6	74.9	78.1
baseline+	✓	✓	✓	✓	✓				81.6	75.0	78.2
Mask TTD	✓	✓	✓	✓	✓	✓			79.0	79.7	79.4
Mask TTD	✓	✓	✓	✓	✓	✓	✓		53.9	65.4	53.7
Mask TTD	✓	✓	✓	✓	✓	✓		✓	37.8	25.6	30.5

ating results are from the mask branch with the same mask grouping method introduced in the following section, and both models share the same training dataset, hyper-parameters, and training mechanism. The detection performance is evaluated using the improved and more reasonable evaluating method in [5], which computes the exact IoU between the polygons. The results in Table III showed the proposed Mask TTD can significantly outperforms the baseline method by a large margin. The speed of final Mask TTD is 1.6 FPS. If we used off-the-shelf functions instead of proposed polygonal grouping method, the running time is 1.4 FPS. In addition, the results showed that only using mutual promotion (Mask TTD without tightness prior and frontier mask) can result in significant improvement. Other components are also quantitatively demonstrated to be effective for the final performance. Moreover, we conducted experiments to test the impact of the labeling sequence on direct regression methods in Mask TTD. The results in Table III indicate that a weak point sequential protocol (using the minimum value of x as the first point and proceeding clockwise) would significantly degrade the performance, and strong sequential interference (using the minimum value of y as the first point) would further degrade the results. Therefore, using the reading direction annotations suggested in Section III is important for training the Mask TTD.

In addition, for curved text, an IoU threshold of 0.5 may not be sufficiently tight to detect the curved text; hence, we also present the detection performance under a stricter IoU threshold (0.6, 0.7, 0.8) in Table IV. It can be seen that Mask TTD significantly outperforms the existing method [5], especially under an IoU threshold of 0.8 (16% higher in terms of F-measure).

With respect to Total-text, because the number of annotated points is not fixed, this dataset cannot be trained directly with Mask TTD. We followed a similar method as [5] in combination with the reading sequence discussed in Section III to complement the number of points to 14 (re-sampling the points in the original boundary). The results of the Total-text test set are provided in Table II. Note that for fair comparison, we utilized an existing evaluation protocol [6] to evaluate our result. The results show that Mask TTD outperforms [6] by a large margin, indicating its effectiveness in curved text detection.

TABLE IV

EXPERIMENTS ON CTW1500 BASED ON HIGH IOU THRESHOLDS

Methods	IoU	R (%)	P (%)	H (%)
CTD+TLOC [5]	0.6	61.3	67.7	64.3
Mask TTD	0.6	74.5	68.4	71.3
CTD+TLOC [5]	0.7	44.4	49.1	46.6
Mask TTD	0.7	62.2	57.1	59.5
CTD+TLOC [5]	0.8	18.6	20.6	19.5
Mask TTD	0.8	36.9	33.9	35.3

The results of the CUTE80 test set (80 images) are included in the Table II. Because many small but distinguishable text items are unlabeled, the detection results of other recognizable text produced by our method would be expected to significantly degrade performance. Therefore, the results are presented by regarding those that are obviously recognizable but are unlabeled as "not care" regions. It is worth mentioning that other researcher [7] evaluated its performance by computing the intersection between the rotated rectangles (transferring polygonal ground truths to the rotated rectangles), whereas our method was evaluated by the strict polygonal intersections. Nonetheless, our method still outperforms theirs [7] by a large margin.

Selected detection results of the three datasets are shown in Figure 7, which demonstrates the robust ability of Mask TTD to detect various types of scene text. A few failures are also exhibited in the last column of this figure, and they might be caused by the limited availability of curved data. Another reason for these failures may be the difficulty associated with the separation of stacked texts without semantic information.

B. Experiments on Popular Non-Curved Datasets

Dataset - ICDAR 2017 RCTW. RCTW-17 [12], which is the largest Chinese-based dataset, includes 8,285 training images and 4,229 testing images. The dataset mainly consists of natural images captured by smartphone cameras, whereas the other images have digital origins and are mainly taken from screen-shots. All images are carefully annotated, and because of the scene variance and high resolution of phone images, many regions that contain small text prevail, which increases the complexity of this dataset. To preserve the small text information, we simply train all the images with a size

TABLE V

EXPERIMENTAL RESULTS ON LARGE-SCALE CHINESE DATASET RCTW-17. THE RESULTS ARE FROM THE RCTW-17 OFFICIAL COMPETITION WEBSITE (BEFORE 01/31/2018). THE MAJOR RANKING METRIC IS AP

Algorithm	**(AP %)**	R (%)	P (%)	H (%)
Organized_team [12]	35.9	40.4	76.0	52.8
SCUT_MBCNN [12]	49.4	51.8	73.6	60.8
ocr [12]	51.4	52.3	68.3	59.2
Argman [12]	51.8	60.0	73.4	66.1
EasyOCR [12]	53.2	57.0	**78.4**	66.0
unhzyx [12]	53.5	58.7	76.8	66.5
IVA [12]	55.5	55.2	66.1	60.2
Sensetime [12]	55.5	57.8	70.6	63.6
gmh [12]	55.5	57.8	70.6	63.6
NLPR_PAL [12]	56.0	57.3	77.2	65.8
Foo & Bar [12]	62.3	58.5	75.5	65.9
Mask TTD	**68.4**	**64.3**	77.2	**70.2**

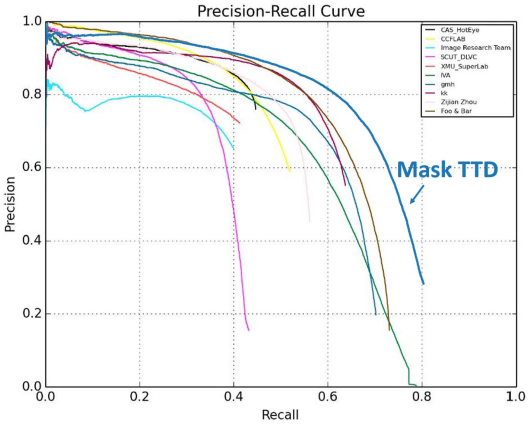


Fig. 8. PR curve figure of RCTW-17 results. Our Mask TTD shows superior performance, especially in terms of recall ability.

of 1600x1800. As shown before [12], this dataset takes AP as the primary metric and ranks submissions accordingly. Basically, mAP is the normalized area below the precision-recall (PR) curve, which is adopted by PASCOL VOC [61]. As text is the only foreground category in this competition, the metric is AP. Because AP is highly relevant to the recall rate, we used a threshold of 0.3 PNMS. Additionally, we used a 0.01 confidence threshold to preserve most of the detection results. To train our method, we used an existing interpolating method [5] to complement the 14 points for all the training bounding boxes, and used the provided training set without any data augmentation, online hard example mining (OHEM) [62] or specific post-processing.

The detection results are presented in Table V (we used a rectangular grouping method to test the results). The results indicate that Mask TTD can achieve 68.4% AP, and 70.2% Hmean, outperforming all previous methods by a large margin. The PR curve is shown in Figure 8 and enables to clearly visualize the superior performance of the proposed Mask TTD. Some of the detection results are shown in Figure 9.

Dataset - MSRATD500. MSRA-TD500 was introduced by [11]. It contains 500 images with multi-oriented English and Chinese scene text. We used the model trained

Fig. 9. Detection results of RCTW-17 dataset. The last figure of the second row shows the results with a very low confidence threshold.

TABLE VI

EXPERIMENTAL RESULTS ON MSRA-TD500

Algorithm	R (%)	P (%)	H (%)
Yao et al. [11]	63.0	63.0	60.0
Zhang et al. [31]	67.0	83.0	74.0
RRPN [35]	68.0	82.0	74.0
He el al. [40]	70.0	77.0	74.0
Yao et al. [63]	75.3	76.5	75.9
EAST [39]	67.4	87.3	76.1
SegLink [2]	70.0	86.0	77.0
Wu et al. [8]	78.0	77.0	77.0
PixelLink [42]	73.2	83.0	77.8
FSTN [41]	77.1	**87.6**	82.0
Mask TTD	**81.1**	85.7	**83.3**

by RCTW-17 directly to test this dataset without using its original training data. The results on this dataset are provided in Table VI along with those of other state-of-the-art methods. It is shown that our method outperforms all previous state-of-the-art approaches, especially in terms of the recall rate. It is worth mentioning that our results are evaluated by the exact IoU between oriented rectangles by [5] instead of using the approximate protocol of [11]; otherwise, the results can be improved further. Some visualization results are shown in Figure 10, which indicates that our method can tightly and accurately recall multi-oriented or very small text. Moreover, as shown by the image at the bottom of the middle column of Figure 10, some mirrored but recognizable text is detected by Mask TTD, but this text is not annotated, which may degrade the accuracy of our method to a certain extent.

Dataset - ICDAR 2015 Competition Challenge 4 Incidental Scene Text. The ICDAR2015 - Incidental Scene Text dataset includes 1,000 training images and 500 testing images. This dataset focuses on an incidental scene where

Fig. 10. Detection results of MSRA-TD500 dataset.

TABLE VII
EXPERIMENTAL RESULTS ON THE ICDAR2015 CHALLENGE 4 TASK 1

Algorithm	R (%)	P (%)	H (%)
SegLink [41]	76.5	74.7	75.6
SSTD [64]	73.9	80.2	76.9
WordSup [3]	77.0	79.3	78.2
RRPN [35]	77.1	83.5	80.2
EAST [41]	78.3	83.3	80.7
NLPR-CASIA [40]	80.0	82.0	81.0
R2CNN [36]	79.7	85.6	82.5
PixelLink[42]	82.0	85.5	83.7
FTSN [41]	80.1	**88.7**	84.1
SLPR [43]	83.6	85.5	84.5
Mask TTD rect	**87.6**	86.6	**87.1**
Mask TTD rect (0.7 IoU)	66.6	71.9	69.1
Mask TTD quad (0.7 IoU)	68.7	74.1	71.3
Mask TTD rect (0.8 IoU)	36.6	39.4	37.9
Mask TTD quad (0.8 IoU)	41.4	44.6	42.9

text may appear in any orientation and at any location with small size or low resolution. The annotations of the ICDAR 2015 ground truth are marked at the word level.

Because of the limited size of the training set, many previous state-of-the-art methods adopt data augmentation to achieve good results. For example, the use of SynthText [59] data to pretrain the model or generating pseudo samples from images with multi-scale sizes or captured at rotated angles can significantly improve the performance on this dataset. Additionally, using OHEM to recycle the hard training data is also demonstrated to be effective to improve the results. Because the data augmentation strategy various methods to method, it is difficult to conduct a strictly fair comparison. Therefore, we used similar approaches and data augmentation to train and test our method. We randomly scaled the image short size to [720, 960, 1200, 1600]. In addition, images were rotated through $[-10°, 10°]$, but we did not utilize OHEM. For training, we utilized our final RCTW-17 model as the pretrained model, and we used SynthText data to finetune the model so that it could detect word level English character. Then, we used official ICDAR 2015 training images to further finetune our model for this dataset. Moreover, some false positives with very low recognition confidence were simply suppressed to improve the precision. The results of this dataset are shown in Table VII. The results demonstrate that the Mask TTD can achieve state-of-the-art performance on this dataset.

Fig. 11. Detection results of ICDAR 2015 challenge 4 dataset. From left to right are: Mask results, rectangular grouping results, and quadrilateral grouping results.

Moreover, we tested our method with quadrilateral or rectangular grouping methods under stricter IoU thresholds (0.7 and 0.8) and found that the quadrilateral results are obviously more accurate than the rectangular results under these conditions. Examples with these grouping methods are shown in Figure 11.

V. CONCLUSION

The detection of arbitrarily shaped scene text remains a challenge, especially with regard to curved text. In this study, we presented Mask TTD, a simple but effective framework for the detection of scene text with arbitrary shapes. We analyzed the advantages and disadvantages of both direct regression and FCN-based methods. Ultimately, we combined the advantages of both by using a conceptually simple branched mutual promotion method to integrate polygonal boundary regression and text mask prediction to significantly improve text detection performance. Moreover, by using a tightness prior and text frontier learning, we successfully enhanced the mask prediction. In addition, we examined the reason for the direct regression method being highly sensitive to the labeling sequence, and we conducted experiments to support our analysis.

Our experiments on all existing datasets containing curved text showed that the proposed Mask TTD can achieve the best performance, even under stricter IoU thresholds. On the well-known ICADAR 2015 challenge 4 dataset, MSRA-TD500, and the largest Chinese dataset RCTW-17, Mask TTD was shown to remarkably outperform all state-of-the-art methods by a large margin.

On the basis of our analysis of samples that produced an error, we concluded that most false positives are caused by ambiguous transverse and longitudinal stacked text or box-in-box text. This suggested that heuristic post-processing or the NLP method may be suitable for improving the precision, which is a topic for further research. As our method could be used to generate common datasets effectively, we expect it to be possible to use common datasets to evaluate polygonal results in the future.

REFERENCES

[1] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. ECCV*, 2016, pp. 56–72.

[2] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *Proc. CVPR*, Jun. 2017, pp. 2550–2558.

[3] H. Hu, C. Zhang, Y. Luo, Y. Wang, J. Han, and E. Ding, "WordSup: Exploiting word annotations for character based text detection," in *Proc. ICCV*, Oct. 2017, pp. 4940–4949.

[4] X.-C. Yin, W.-Y. Pei, J. Zhang, and H.-W. Hao, "Multi-orientation scene text detection with adaptive clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1930–1937, Sep. 2015.

[5] L. Yuliang, J. Lianwen, Z. Shuaitao, and Z. Sheng, "Detecting curve text in the wild: New dataset and new solution," 2017, *arXiv:1712.02170*. [Online]. Available: https://arxiv.org/abs/1712.02170

[6] C. K. Ch'ng and C. S. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in *Proc. ICDAR*, Nov. 2017, pp. 935–942.

[7] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Syst. Appl.*, vol. 41, no. 18, pp. 8027–8048, 2014.

[8] Y. Wu and P. Natarajan, "Self-organized text detection with minimal post-processing via border learning," in *Proc. CVPR*, Oct. 2017, pp. 5010–5019.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, May 2015, pp. 3431–3440.

[10] D. Karatzas *et al.*, "ICDAR 2015 competition on robust reading," in *Proc. ICDAR*, Aug. 2015, pp. 1156–1160.

[11] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. CVPR*, Jun. 2012, pp. 1083–1090.

[12] B. Shi *et al.*, "ICDAR2017 competition on reading Chinese text in the wild (RCTW-17)," in *Proc. ICDAR*, Nov. 2017, pp. 1429–1434.

[13] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions," in *Proc. ICDAR*, Aug. 2003, pp. 682–687.

[14] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," in *Proc. CVPR*, vol. 2, Jun./Jul. 2004, p. 2.

[15] M. R. Lyu, J. Song, and M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 2, pp. 243–255, Feb. 2005.

[16] Y. Liu, S. Goto, and T. Ikenaga, "A contour-based robust algorithm for text detection in color images," *IEICE Trans. Inf. Syst.*, vol. E89-D, no. 3, pp. 1221–1230, Mar. 2006.

[17] K. Wang and S. Belongie, "Word spotting in the wild," in *Proc. ECCV*, 2010, pp. 591–604.

[18] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. CVPR*, Jun. 2010, pp. 2963–2970.

[19] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Proc. ACCV*, 2010, pp. 770–783.

[20] C. Yi and Y. Tian, "Text string detection from natural scenes by structure-based partition and grouping," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2594–2605, Sep. 2011.

[21] 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.

[22] W. Huang, Z. Lin, J. Yang, and J. Wang, "Text localization in natural images using stroke feature transform and text covariance descriptors," in *Proc. ICCV*, Dec. 2013, pp. 1241–1248.

[23] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced MSER trees," in *Proc. ECCV*, 2014, pp. 497–511.

[24] L. Kang, Y. Li, and D. Doermann, "Orientation robust text line detection in natural images," in *Proc. CVPR*, Jun. 2014, pp. 4034–4041.

[25] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 970–983, May 2014.

[26] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. L. Tan, "Text flow: A unified text detection system in natural scene images," in *Proc. ICCV*, Dec. 2015, pp. 4651–4659.

[27] 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.

[28] 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.

[29] Z. Zhong, L. Jin, S. Zhang, and Z. Feng, "DeepText: A unified framework for text proposal generation and text detection in natural images," 2016, *arXiv:1605.07314*. [Online]. Available: https://arxiv.org/abs/1605.07314

[30] H. Cho, M. Sung, and B. Jun, "Canny text detector: Fast and robust scene text localization algorithm," in *Proc. CVPR*, Jun. 2016, pp. 3566–3573.

[31] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," in *Proc. CVPR*, Jun. 2016, pp. 4159–4167.

[32] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: A fast text detector with a single deep neural network," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4161–4167.

[33] S. Zhang, Y. Liu, L. Jin, and C. Luo, "Feature enhancement network: A refined scene text detector," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 2612–2619.

[34] Z. Zhong, L. Sun, and Q. Huo, "Improved localization accuracy by LocNet for faster R-CNN based text detection," in *Proc. ICDAR*, Nov. 2017, pp. 923–928.

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

[36] Y. Jiang *et al.*, "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*. [Online]. Available: https://arxiv.org/abs/1706.09579

[37] H. Li, P. Wang, and C. Shen, "Towards end-to-end text spotting with convolutional recurrent neural networks," in *Proc. ICCV*, Jun. 2017, pp. 5238–5246.

[38] Y. Liu and L. Jin, "Deep matching prior network: Toward tighter multi-oriented text detection," in *Proc. CVPR*, Jun. 2017, pp. 1962–1969.

[39] X. Zhou *et al.*, "EAST: An efficient and accurate scene text detector," in *Proc. CVPR*, Jun. 2017, pp. 5551–5560.

[40] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," in *Proc. ICCV*, Jun. 2017, pp. 745–753.

[41] Y. Dai *et al.*, "Fused text segmentation networks for multi-oriented scene text detection," in *Proc. ICPR*, Aug. 2018, pp. 3604–3609.

[42] D. Deng, H. Liu, X. Li, and D. Cai, "PixelLink: Detecting scene text via instance segmentation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 6773–6780.

[43] Y. Zhu and J. Du, "Sliding line point regression for shape robust scene text detection," 2018, *arXiv:1801.09969*. [Online]. Available: https://arxiv.org/abs/1801.09969

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

[45] A. Shahab, F. Shafait, and A. Dengel, "ICDAR 2011 robust reading competition challenge 2: Reading text in scene images," in *Proc. ICDAR*, Sep. 2011, pp. 1491–1496.

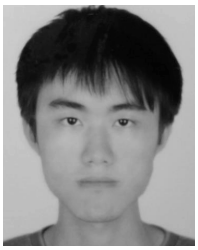
[46] D. Karatzas *et al.*, "ICDAR 2013 robust reading competition," in *Proc. ICDAR*, Aug. 2013, pp. 1484–1493.

[47] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "COCO-text: Dataset and benchmark for text detection and recognition in natural images," 2016, *arXiv:1601.07140*. [Online]. Available: https://arxiv.org/abs/1601.07140

[48] R. Nagy, A. Dicker, and K. Meyer-Wegener, "NEOCR: A configurable dataset for natural image text recognition," in *Proc. Int. Workshop Camera-Based Document Anal. Recognit.*, 2011, pp. 150–163.

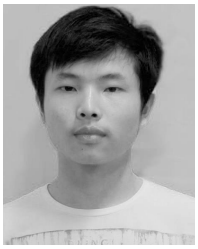
[49] M. Iwamura, N. Morimoto, K. Tainaka, D. Bazazian, L. Gomez, and D. Karatzas, "ICDAR2017 robust reading challenge on omnidirectional video," in *Proc. ICDAR*, vol. 1, Nov. 2017, pp. 1448–1453.

[50] N. Nayef *et al.*, "ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification—RRC-MLT," in *Proc. ICDAR*, vol. 1, Nov. 2017, pp. 1454–1459.
[51] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, Jun. 2017, pp. 2961–2969.
[52] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. NIPS*, 2015, pp. 2017–2025.
[53] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, Jun. 2015, pp. 1440–1448.
[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, Jun. 2009, pp. 248–255.
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.
[56] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2016, *arXiv:1612.03144*. [Online]. Available: https://arxiv.org/abs/1612.03144
[57] D. Eppstein, M. Overmars, G. Rote, and G. Woeginger, "Finding minimum area*k*-gons," *Discrete Comput. Geometry*, vol. 7, no. 1, pp. 45–58, 1992.
[58] T. Chen *et al.*, "MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems," 2015, *arXiv:1512.01274*. [Online]. Available: https://arxiv.org/abs/1512.01274
[59] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. CVPR*, Jun. 2016, pp. 2315–2324.
[60] W. Wang *et al.* (2017). *An MXNet Implementation of Mask R-CNN*. [Online]. Available: https://github.com/TuSimple/mx-maskrcnn
[61] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.
[62] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. CVPR*, Jun. 2016, pp. 761–769.
[63] 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: https://arxiv.org/abs/1606.09002
[64] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in *Proc. ICCV*, Jun. 2017, pp. 3047–3055.



Yuliang Liu received the B.S. degree in electronic and information engineering from the South China University of Technology, Guangdong, China, in 2016. He is currently pursuing the Ph.D. degree with the Deep Learning and Vision Computing Lab (DLVClab), South China University of Technology, under the supervision of Prof. L. Jin. He works on scene text understanding, handwritten character recognition, document analysis, and deep learning-based text detection and recognition. He has actively participated in several international conferences and scientific competitions, and his team won the champion of the competition for ICDAR 2017 multi-lingual scene text detection task and end-to-end scene text detection and classification task.

Lianwen Jin received the B.S. degree from the University of Science and Technology of China, Anhui, China, in 1991, and the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 1996. He is currently a Professor with the School of Electronic and Information Engineering, South China University of Technology. He is the author of more than 100 scientific articles. His research interests include computer vision, optical character recognition, handwriting analysis and recognition, machine learning, deep learning, and intelligent systems. Dr. Jin was a recipient of the award of New Century Excellent Talent Program of MOE in 2006 and the Guangdong Pearl River Distinguished Professor Award in 2011.



Chuanming Fang received the B.S. degree in electronic and information engineering from the Hei Long Jiang University of Science and Technology in 2017. He is currently pursuing the master's degree with the Deep Learning and Vision Calculations Lab (DLVClab), South China University of Technology, Guangdong, China. His research interests include scene text detection, deep learning, and cloud computing.