# OPMP: An Omnidirectional Pyramid Mask Proposal Network for Arbitrary-Shape Scene Text Detection

Sheng Zhang, Yuliang Liu, Lianwen Jin, *Member, IEEE*, Zhongrong Wei, and Chunhua Shen

***Abstract*—Scene text detection methods have achieved significant progresses. However, stack-omnidirectional text dilemma, under-segmentation of very close text words, and over-segmentation of arbitrary-shape long text lines, are still main challenges. Motivated by these problems, we proposed a two stage method called omnidirectional pyramid mask proposal text detector (OPMP). OPMP removes anchor mechanism that requires heuristic non-maximum suppress processing. Instead, it uses an effective pyramid lengthwise and sidewise residual sequence modeling method to produce arbitrary-shape proposals. To accurately extract the features of text shape, OPMP enhances the backbone layers by a multiple arbitrary-shape fitting mechanism. Finally, a multi-grain text classification module is proposed, which reclassifies each text region robustly. Comprehensive ablation studies demonstrate the effectiveness of each proposed component. In addition, experiments on various benchmarks, including ICDAR2015, MLT, MSRA-TD500, CTW1500, and Total-text, show that our method outperforms previous state-of-the-art methods.**

***Index Terms*—Text detection, pyramid sequence modeling, omnidirectional pyramid mask proposal.**

## I. Introduction

SCENE text detection within images and videos is an important component of various intelligent applications based on computer vision techniques [1]–[6]; e.g., image-based geolocation, pilotless automobiles, blind navigation, and multilingual translation. Scene text detection has many challenges, such as a variation in the aspect-ratios and scales, perspective distortions, complex backgrounds, uncontrollable illumination intensity.

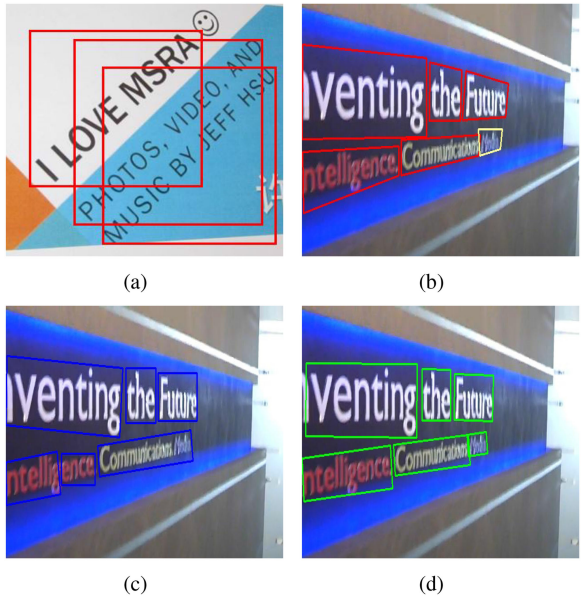



Fig. 1. Challenges for scene text detection: (a) The stack-omnidirectional text dilemma; (b) Ground truths on the ICDAR2015 benchmark; (c) Under/over-segmentation problems in the PixelLink framework [7]; and (d) Our OPMP.

In recent years, omnidirectional and arbitrary-shape text detection methods have attracted much attention [7]–[12]. These methods predict quadrangles or polygons instead of horizontal bounding-boxes to localize the omnidirectional or arbitrary-shape text regions respectively, which remarkably improves the text detection performance for various complex scenes. However, owing to a stack-omnidirectional text dilemma, the performances of state-of-the-art methods are still far from the demands of real-world applications.

- Some methods [6], [13]–[15] rely on the Region Proposal Network (RPN) [13] which requires Non-Maximum Suppression (NMS) [16] procedure to suppress massive redundant proposals. However, such procedure is inappropriate in stack-omnidirectional text cases, as some positive proposals may have mutually suppressed each other. Example is illustrated in Fig. 1(a).
- Although some methods [10], [17] have been proposed to alleviate the issue by introducing rotated proposals, calculating overlaps between numerous quadrilateral proposals is very time-consuming.
- Other methods [7], [18] follow the concept of segmentation to avoid non-maximum suppress; however, these methods suffer from the problems of the under-segmentation of

Manuscript received August 2, 2019; revised January 5, 2020; accepted February 25, 2020. Date of publication March 9, 2020; date of current version December 17, 2020. This work was supported in part by NSFC under Grants 61936003 and GD-NSF 2017A030312006, in part by the National Key Research and Development Program of China under Grant 2016YFB1001405, in part by Guangdong Intellectual Property Office Project 2018-10-1, and in part by GZSTP under Grant 201704020134. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joao M Ascenso. *(Sheng Zhang and Yuliang Liu contributed equally to this work.) (Corresponding author: Lianwen Jin.)*

Sheng Zhang, Lianwen Jin, and Zhongrong Wei are with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: zsscut@sina.com; lianwen.jin@gmail.com; zrwei@foxmail.com).

Yuliang Liu is with The University of Adelaide, Adelaide, SA 5005, Australia. He is now with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: liu.yuliang@mail.scut.edu.cn).

Chunhua Shen is with the School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia (e-mail: chunhua.shen@adelaide.edu.au).

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

Digital Object Identifier 10.1109/TMM.2020.2978630

much close text words and over-segmentation of long text lines, as shown in Fig. 1(c).

To tackle the above issues, we propose a new omnidirectional pyramid mask proposal text detector, namely OPMP. Firstly, in the omnidirectional proposal generation stage, OPMP integrates the novel pyramid lengthwise and sidewise residual sequence modeling (LSRSM) module to solve the under/over-segmentation problems, and generates a suitable number of omnidirectional proposals without the non-maximum suppress procedure, which solves the stack-omnidirectional text dilemma. Secondly, we enhance the feature representation of arbitrary-shape text with the novel multiple arbitrary-shape fitting (MASF) module. In addition, we apply Skip RoIAlign [14] to extract the axis-align and omnidirectional text features separately. Based on the extracted features, the arbitrary-shape mask is further reconstructed by a new multi-grain text classification (MGTC) module, which rescores the final confidence reasonably.

The main contributions of our work are summarized as follows.

1) In proposal generating stage, the proposed OPMP can generate moderate and accurate omnidirectional mask proposals rather than massive redundant proposals from anchor mechanism, and thus OPMP can thoroughly avoid NMS and its negative impact.
2) The proposed pyramid lengthwise and sidewise residual sequence modeling module can effectively solve the problems of the under-segmentation of the adjacent text words and over-segmentation of long text lines.
3) We propose a multiple arbitrary-shape fitting module that can accurately extract the arbitrary-shape text features.
4) The proposed multi-grain text classification can reconstruct the arbitrary-shape text mask precisely, and rescore the final confidence to make the detection result robuster.
5) Our approach achieves state-of-the-art performance on various challenging benchmarks, including ICDAR2015, MLT, MSRA-TD500, CTW1500, and Total-text.

The remainder of the paper is organized as follows. Section II briefly reviews scene text detection. Section III describes the proposed method. Section IV presents a series of experimental results and analyses. Section V presents a concise conclusion summarizing the main points of the work.

## II. Related Work

Various text detection methods have been developed in recent years. We firstly review a few of them in Section II-A, and then compare the proposed OPMP against related works in Section II-B.

### A. Typical Text Detection Works

Typical text detection works can be roughly categorized into hand-crafted and convolutional neural network (CNN)-based works. Prior to the emergence of CNN, text detection methods commonly comprise several steps, such as text component extraction and filtering, component grouping, and candidate filtering. The main step is extracting text components with hand-crafted features. The stroke width transform [19] and maximally stable extremal regions [20] are two typical works for text component extraction. Many preceding methods [21]–[25] have evolved from these two works. Other works of this category are [26], [27]. The most recently proposed methods [28]–[30] use the CNN to detect scene text. Overall, these methods can be roughly classified into regression-based, segmentation-based, and regression/segmentation hybrid works. For the first one, they can be further categorized into two classes according to the grain of regression targets: text word/line-based and text component-based regression methods.

*Text word/line-based regression methods:* Text word/line-based regression methods are mostly motivated by the recent development of general object detection algorithms, where the whole text word/line is viewed as an object to be detected. By applying the text specific inception module and skip RoI-Pooling component, DeepText [15] adapts the Faster R-CNN [13] framework to detect various aspect ratios and scales of scene texts. DMPNet [10] exploits well-designed quadrilateral anchors with different orientations to detect omnidirectional texts. RRPN [17] presents a multi-oriented text detection framework, in which six different orientation anchors are generated at each feature point of the specific feature map. The angle information is a regression target to get more accurate rotated boxes. However, these methods generate massive redundant quadrilateral proposals, which is very time-consuming to conduct the procedure of polygon non-maximum suppress (PNMS) [10], or suffers from the stack-omnidirectional text dilemma. Simultaneously, they are sensitive to the quadrilateral vertex order of ground truths to some extent, especially when using the rotation data-augmentation strategy [7] to train the model. DDR [31] uses a fully convolutional network to directly predict the final quadrilateral from a given point. In the test phase, a multi-scale sliding window strategy is used, which is time-consuming.

*Text component-based regression methods:* It is difficult to apply regression directly to the whole text word/line due to large variations in the text aspect ratios, scales, and orientations. Some methods regress text components and predict the link between them. CTPN [32] is the first method to predict vertically thin text components and then leverage a recurrent neural network to link text components. However, the method only works well for horizontal scene texts. An innovative Markov clustering network was proposed in [33]. This network considers an image as a stochastic flow graph, where the flows are strong between text nodes but weak between remaining nodes. A Markov clustering process is then applied to form text instances from the predicted flow graph. SegLink [34] designed a network by detecting the segments of omnidirectional text and concurrently predicting the linking relationship for 12-neighboring text segments. The final detected texts are output by grouping the segments with the links. However, the method is vulnerable to the unbalanced segment cropping of quadrilateral ground-truths.

*Segmentation-based methods:* Segmentation-based methods classify the whole text image in a pixel-wise text/non-text binarization manner, which is often done by using a fully convolutional network (FCN) [35]. Yao *et al.* [36] suggested predicting both individual characters and the orientation of text

boxes with the FCN in a whole manner. The detection results are then obtained with a grouping process based on three estimated properties of text. In [37], the authors make full use of multi-scale inputs and apply the FCN to predict text blocks. The followings are two CNN forks to predict text lines and do instance-segmentation from the estimated text blocks. Wu *et al.* [38] introduced text, text border, and non-text three-class semantic segmentation, which facilitates the division of neighboring text instances. PixelLink [7] exploited the link prediction between the pixels to conduct text instance segmentation. Although the method works well in some cases, it readily generates false positives for text-like objects, performs poorly in the separation of very adjacent texts, and suffers from long text lines.

*Regression/segmentation-based hybrid methods:* Zhou *et al.* [12] designed the efficient scene text detector EAST, which has two branches; i.e., a segmentation branch predicts the text score map while a regression fork predicts the final box for each point in the text region. Zhong *et al.* [39] developed the anchor-free text detector AF-RPN, which applies both segmentation and regression branches to multiple stages of CNN to acquire omnidirectional proposals for further refinement. However, these methods do not overcome the shortcomings of both regression-based and segmentation-based methods.

### B. Difference With Related Works

*OPMP versus conventional hand-crafted approaches:* Conventional approaches exploit carefully hand-crafted features to extract text components, and employ various heuristic grouping rules to detect text instances. Each subprocess demands well-designed parameters, which leads to imperfect performance and a long train/test runtime of the whole pipeline. The proposed OPMP takes full advantage of the powerful representation ability of CNN to realize excellent text detection performance. Meanwhile, OPMP is fully end-to-end trainable and efficient in the test phase.

*OPMP versus regression-based approaches:* Regression-based approaches generate massive redundant horizontal or quadrilateral proposals, making it time-consuming to conduct the procedure of polygon non-maximum suppression (PNMS) [29] or suffering from the stack-omnidirectional text dilemma. Meanwhile, our experiments show that text component-based regression methods are vulnerable to unbalanced component cropping of the quadrilateral ground-truths, especially for the vertical text instances generated by rotation data-augmentation. The proposed OPMP is a two-stage instance segmentation-based method. It only generates 128 omnidirectional mask proposals empirically for the successive arbitrary-shape text detection stage, and does not adopt time-consuming PNMS during the train phase.

*OPMP versus segmentation-based approaches:* Scene texts have the characteristics of varying sizes, aspect-ratios and shapes, and segmentation-based method is capable to address such challenges. However, segmentation-based approaches are vulnerable to the problems of the under-segmentation of very close text instances and the over-segmentation of arbitrary-shape long text lines. Furthermore, some instance segmentation-based methods have recently been proposed, which make the best use of additional information to improve the division of neighboring texts, such as the text contour, text direction, or linking relationship between adjacent pixels. However, the above methods are almost all single-stage, single-output-level, pixel-wise classification segmentation methods. They do not take full advantage of the whole text region classification, and still face the above under/over-segmentation problems. Our OPMP is a two-stage instance segmentation approach with omnidirectional pyramid mask proposal, which simultaneously integrates the three novel components of pyramid LSRSM, MASF, and MGTC to deal with the above problems effectively.

## III. Methodology

### A. Overall Architecture

The overall architecture of our OPMP is elaborated in Fig. 2 and includes three parts: the network backbone design, omnidirectional pyramid mask proposal, and arbitrary-shape text detection. The network backbone for text feature extraction takes the union of ResNet-50 [40] and feature pyramid network (FPN) [41]. FPN is the most effective feature extraction network for various object detection tasks, as it can exploit the inherent multi-scale, pyramidal hierarchy of deep convolutional networks to construct feature pyramids with marginal extra cost. Specifically, for the ResNet-50 subnetwork, we simply remove the final full connection layer. Meanwhile, we adopt the group normalization (GN) [42] instead of the batch normalization [43] regularization strategy to accelerate our model training, because group normalization is insensitive and robust over a wide range of batch sizes. The omnidirectional pyramid mask proposal stage involves a pyramid module, two branches of the pixel-wise text/non-text classification, and the predictions of links between adjacent pixels, which roughly recalls various omnidirectional texts. The last arbitrary-shape text detection stage involves multiple arbitrary-shape fitting module and a multi-grain text classification module, which is designed to detect the arbitrary-shape texts robustly.

### B. Omnidirectional Pyramid Mask Proposal

As is well-known, Recurrent Neural Networks (RNN) [44] can strengthen or weaken the relationship among broad space sequence features, which is determined by the inherent strong control of space sequence information flow of the recurrent component. Assuming that the ground truths are words, the recurrent component prevents information flow between adjacent words via the supervision information of blank background between words. In contrast, assuming that the ground truths are text lines, the information flow among long scope context is enhanced by selectively overlooking the blank spaces between adjacent words. Theoretically, we can alleviate the problems of the under-segmentation of much close text words and the over-segmentation of long text lines with RNN. In fact, employing RNN to perceive long scope context information has been shown to be effective and robust by text detector CTPN [32]. However, CTPN only builds the empirically horizontal linking,

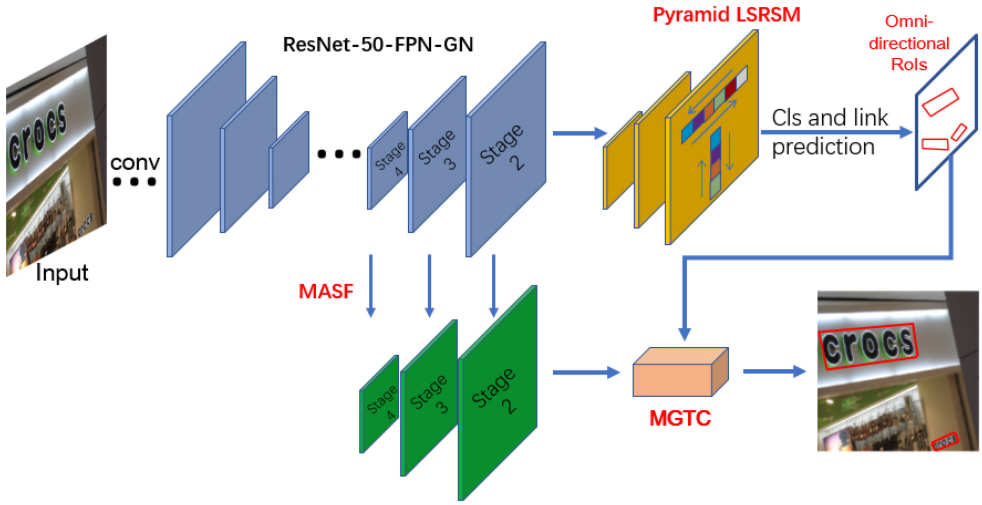



Fig. 2. Overall architecture of our end-to-end trainable text detector OPMP. PLSRSM: Pyramid lengthwise and sidewise residual sequence modeling. MASF: Multiple arbitrary-shape fitting. MGTC: Multi-grain text classification.

and leverages the length of the whole row of a specific feature map as input time steps of RNN, which is not efficient owing to numerous time steps and the serialization computation in RNN. Additionally, to guarantee a constant time steps for the recurrent component, all input images must be of the same fixed size, which means that CTPN is merely trainable in a single-scale manner, giving up the advantage of multi-scale training that improves model performance.

To solve the under/over-segmentation problems effectively, and make the model trainable in a multi-scale way, we propose a pyramid lengthwise and sidewise residual sequence modeling module with a classic LSTM component, which concurrently generates a suitable number of omnidirectional text mask proposals to solve the stack-omnidirectional text dilemma.

*1) Pyramid Lengthwise and Sidewise Residual Sequence Modeling (PLSRSM):* We devise a left-to-right, right-to-left, top-to-bottom, bottom-to-top, pyramid lengthwise and sidewise residual sequence modeling module, which is obviously different from [45]. Firstly, in Fig. 3, we reduce the last feature maps of $Stage\,2$ with channel dimension $C_0$ to feature maps $F_d$ with dimension $C$ ($C = \frac{1}{4}C_0$) via $1 \times 1$ convolution. For the horizontal bi-direction, we then reshape the dimension-reduced feature maps $F_d$ with shape $(C, H_f, W_f)$ into feature maps $F_r$ with shape $(C, H_f \times W_f/TS, TS)$. Secondly, we transpose the feature maps $F_r$ into feature maps $F_t$ with shape $(TS, H_f \times W_f/TS, C)$. To consider the model efficiency and effectiveness as well as the text width characteristic simultaneously, $TS$ which is the time steps of horizontal LSTM modules, is set at 16 empirically for text words and 32 for text lines, since horizontal texts have larger aspect-ratios of width/height than vertical texts, and the widths $W_f$ of feature maps must be divided by $TS$ exactly with no remainder for feature alignment. Thirdly, we model the large-scale space sequence information via LSTM component with feature maps $F_t$. As for vertical bi-direction, we initially transpose the corresponding dimension-reduced feature maps $F_d$ in the height and width dimensions, and then repeat the above horizontal bi-direction operations except $TS$

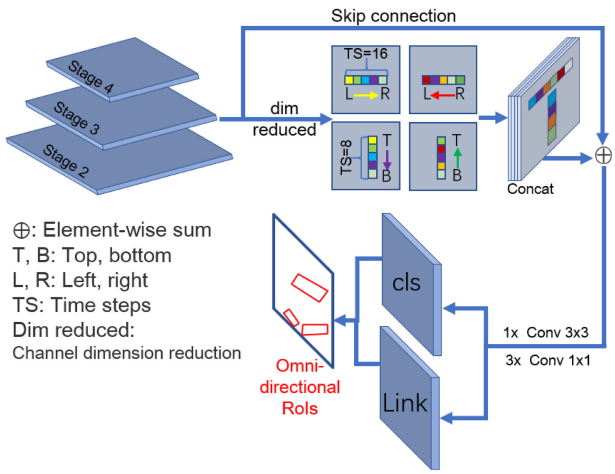



Fig. 3. The elaboration of the pyramid lengthwise and sidewise residual sequence modeling (PLSRSM) module.

is respectively 8 and 16 for text words and lines empirically. In fact, the setting of $TS$ in the vertical direction is not only for the similar reasons of horizontal texts, but also there are fewer steeply inclined or vertical texts than horizontal texts, which leads less vertical context information to be needed in terms of most texts. Besides, smaller $TS$ can accelerate the sequence modeling in the vertical bi-directions. Finally, to restore original space locations of all features, we conduct all above operations inversely with the outputs of LSTM modules, and concatenate them along the channel dimension to ascend dimension to $C_0$. To quickly model large-scale context information and prevent the model from over-fitting, inspired by well-known residual learning [40], we add the concatenated feature maps produced by the lengthwise and sidewise sequence modeling to the last feature maps of $Stage\,2$, fuse them with four successive convolutions in Fig. 3, and finally obtain lengthwise and sidewise residual sequence modeling (LSRSM) module. Fig. 3 clearly shows that our LSRSM module is much more flexible and specific for

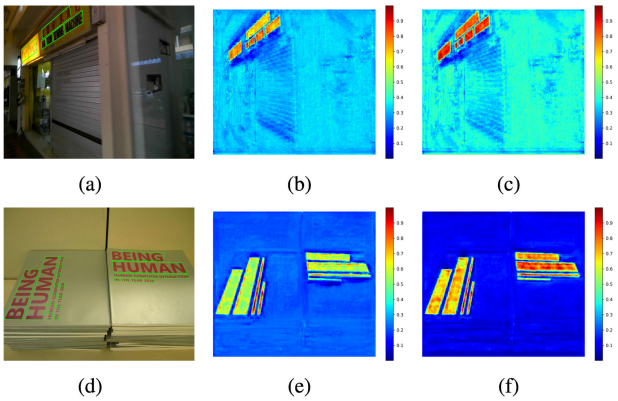



Fig. 4. Effectiveness of our pyramid lengthwise and sidewise residual sequence modeling module. Figures in the first column are input images with text ground truths (green quadrilaterals). Second and third columns are feature maps separately extracted from the input and output of the LSRSM module corresponding to the $Stage$ 2.

omnidirectional and arbitrary-shape texts, since LSRSM does not model the space sequential features with the whole row $W_f$ or column $H_f$ features as time steps; the time steps $TS$ are merely parts of the whole row or column features, which provides several advantages.

- The model can be trained in a fully end-to-end multi-scale fashion to improve the performance, and does not require that all the input images are of the same fixed size, in contrast with the case for the CTPN [32].
- The model can perform much faster because (a) the number of channel dimensions is substantially reduced (from 128 to 32) before each modeling of space sequential features, resulting in 32 input and output feature dimensions for all full connection operations in LSTM; (b) fewer time steps are used in LSRSM module to alleviate the serial computation shortcoming of LSTM, and (c) NVIDIA TITAN X devices have a strong parallel computation capability with the tremendous batch size resulted by the reshape operation in LSRSM.

To further solve the under/over-segmentation problems, we continue to apply the LSRSM module to two other scale feature maps of $\{Stage\ 3, Stage\ 4\}$ in Fig. 3, forming the pyramid lengthwise and sidewise residual sequence modeling module PLSRSM. In the LSRSM modules of $\{Stage\ 3, Stage\ 4\}$, the values for $TS$ are $\{\frac{1}{2}, \frac{1}{4}\}$ of the values of $TS$ in the LSRSM module of $Stage\ 2$, which is efficient and effective. The feature maps extracted from the input and output of PLSRSM module are visualized in Fig. 4, obviously, the difference values between foreground and background feature values of very close text words are enlarged, by contrary, the background features between different words of each long text line are weakened to strengthen the relationship, which demonstrates the effectiveness of our PLSRSM module for solving the under/over-segmentation problems in Fig. 1.

After the operation of PLSRSM module for the large scale context feature modeling, we predict two branches with $1 \times 1$ convolution in each pyramid level, one being the pixel-wise text/non-text classification branch, and the other being the fork

**Algorithm 1:** Omnidirectional Text Proposals Generation

**Input**: Positive text pixel score map $MP$ and 8-neighbour link score maps $ML$.
**Parameter**: Positive pixel and link threshold $p\_th = 0.8$, $l\_th = 0.7$.
**Output**: Positive text pixels $P$ and links set $L$, arbitrary-shape text mask proposals $\mho$, omnidirectional text proposals $Q$.

Let $\mho = \emptyset, Q = \emptyset, RO = \emptyset, mask = \emptyset, idx = 0$.
$P = thres(MP, p\_th)$, $L = thres(ML, l\_th)$, $points = where(P > 0)$, $groups = map(points, -1)$
**For** $point \in points$ **do**
  **if** $groups(point)\ is\ -1$ **then**
    $N = get\_neighbours\_8(point)$.
    **while** $n \in N$ **do**
      **if** $P(n) = P(point)\ \&\ L(n) > 0$ **then**
        $groups(\cdots groups(point)) = n$.
**For** $point \in points$ **do**
  $ro = groups(\cdots groups(point))$
  **if** $ro \notin RO$ **then**
    $RO(ro) = idx + 1, mask(point) = idx + 1$.
  **else**
    $mask(point) = RO(ro)$
**For** $i \in range(1, max(mask) + 1)$ **do**
  $\mho = \mho \bigcup \{cv2.findContours(mask = i)\}$
**For** $i \in \mho$ **do**
  $Q = Q \bigcup \{cv2.minAreaRect(i)\}$
**return** omnidirectional text proposals Q.

of the link predictions for 8-neighbour pixel pairs. We then group all text pixels with all pixel-pair links to different text instances by Algorithm 1, where each omnidirectional proposal is an external minimum area quadrangle of the arbitrary-shape text mask proposal. Only a suitable number of omnidirectional pyramid mask proposals are produced, eliminating the need for the classic non-maximum suppress algorithm that suppresses redundant proposals, and overcoming the stack-omnidirectional text dilemma in Fig. 1(a).

### C. Arbitrary-Shape Text Detection

For the sake of the under-segmentation of very close text words, we adopt similar shrunk polygon labels [12] to supervise the proposal subnetwork training, which leads to the generation of fine-grained omnidirectional pyramid mask proposals. Additionally, instance segmentation is limited to the receptive field of the single pixel and ignores the global information of the whole text region, and instance segmentation is thus vulnerable to the misclassification of text-like objects and noise disturbance, such as blurring, occlusion, and high exposure.

*1) Multiple Arbitrary-Shape Fitting (MASF):* To model various text characteristics accurately in the arbitrary-shape text detection stage, we propose the multiple arbitrary-shape fitting module, which is inspired by deformable convolution [46]. Specifically, deformable convolution has two branches, referred

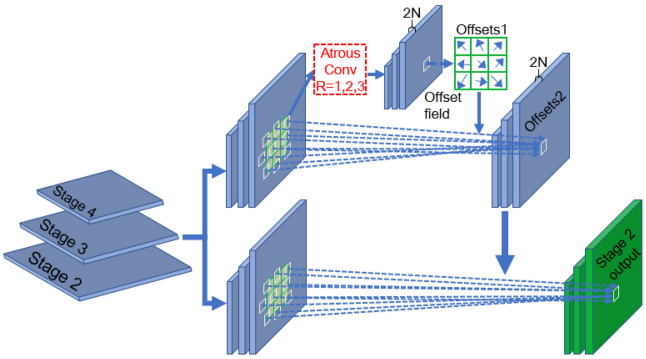



Fig. 5. The elaboration of the multiple arbitrary-shape fitting (MASF) module. $N$ stands for the number of convolution kernel elements, $R$ is dilation rate of the atrous convolution [47].

to Fig. 5, one being the offset prediction of original sampling points to calculate the new sampling points of features for each convolution kernel, and the other being the common convolution with newly sampled features for each convolution kernel. However, the $3 \times 3$ common convolutions in the offset prediction branches have the limited receptive fields. Simultaneously, different atrous convolutions [47] of variable receptive fields not only have larger receptive fields for the rapid perception of context information but also solve the multi-scale problem of semantic segmentation effectively. We thus replace the $3 \times 3$ common convolutions with different atrous convolutions in the offset prediction branches of deformable convolutions to establish three improved deformable convolutions. Furthermore, both $3 \times 3$ common convolutions and atrous convolutions are **regular-shape** convolutions while the offsets of the feature point positions for arbitrary-shape texts are **irregular**. We recursively apply the improved deformable convolutions in the offset prediction branches of three original deformable convolutions [46], which fit the arbitrary-shape text robustly, as the MASF module shows in Fig. 5. Finally, three multi-improved deformable convolutions are further applied on the levels of different network stages separately, creating an effective MASF module that extracts arbitrary-shape text features perfectly for the following multi-grain text classification module. Fig. 6 shows that our MASF module can enhance the feature representation of arbitrary-shape texts perfectly.

*2) Multi-Grain Text Classification (MGTC):* In the top row of Fig. 8, considering to classify each text with two grain-sizes, Fig. 8(a) is to acquire the score for each text region by averaging the predicted text pixel scores, Fig. 8(b) is to extract each text feature by RoIPooling [13] or RoIAlign [14], and further classify it **wholly** with convolution and full connection operations. Obviously, Fig. 8(a) is limited to the receptive field of single pixel, and ignores the global information of the whole text, which is vulnerable to the misclassification of text-like objects and noise disturbance, such as blurring, occlusion, high exposure etc. Fig. 8(b) is constrained by the text shape, such as curved texts, which will extract much background information and produce a low text score. Therefore, we propose a multi-grain text classification module following the MASF module, which takes full advantages of different grain-size classification methods.

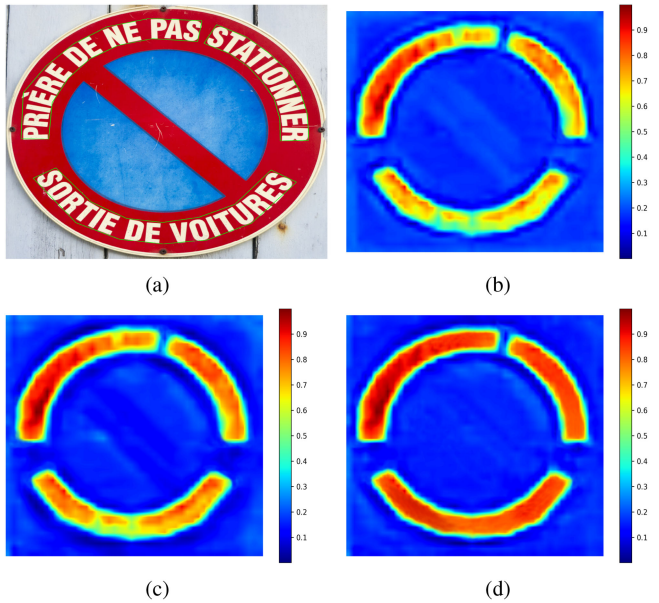



Fig. 6. (a) Input image with green ground truth labels; (b) common convolution; (c) deformable convolution [46]; (d) multi-improved deformable convolution in our MASF module.

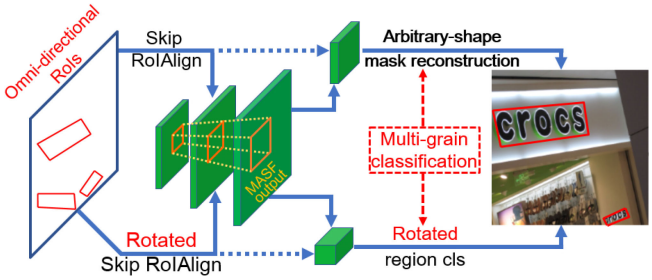



Fig. 7. The elaboration of the multi-grain text classification (MGTC) module.

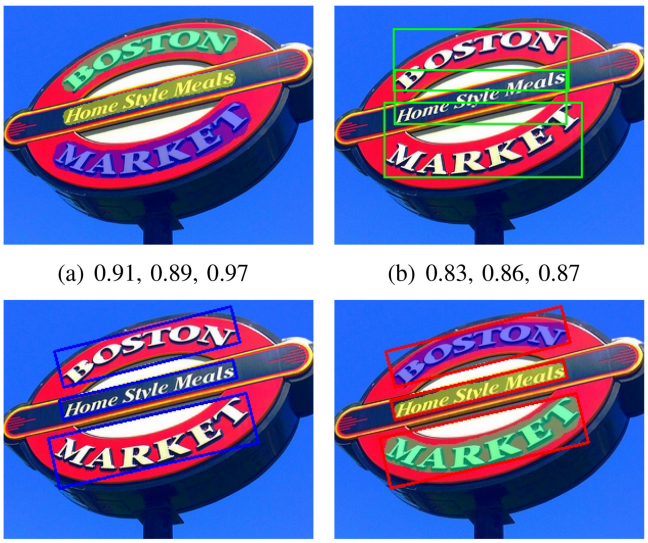
Concretely, it comprises two branches. One is the coarse-grained omnidirectional text classification **wholly**, and the other is the fine-grained arbitrary-shape text mask reconstruction.

For the coarse-grained arbitrary-shape text classification, we firstly expand the sizes of all omnidirectional proposals $Q$ by Equations 1 and 2.

$$min\_sides = \min(W_q, H_q) \tag{1}$$

$$\tilde{Q} = (0\ 0\ s_0\ s_0\ 0) \odot min\_sides + (X_{cq}\ Y_{cq}\ W_q\ H_q\ \theta_q), \quad q \in Q \tag{2}$$

where $\odot$ is the element-wise multiplication operation, $\{X_{cq}\ Y_{cq}\ W_q\ H_q\ \theta_q\}$ are the center coordinates, widths, heights, angles for all omnidirectional proposals $Q$, $s_0 \in [0.07, 0.2]$ is an empirical proposal scaling factor and different benchmarks have specific values. Then, to extract less background feature information and escape from extracting adjacent text features, we improve the feature extraction of each whole text by Rotated Skip RoIAlign with $\tilde{Q}$ in Fig. 8(c), which is obviously tighter and better than the Skip RoIAlign [48] used in Fig. 8(b). Finally,




(a) 0.91, 0.89, 0.97 (b) 0.83, 0.86, 0.87

(c) 0.87, 1.00, 0.93 (d) 0.90, 0.93, 0.96

Fig. 8. Comparison of different grain-size text classifications on CTW1500: (a) fine-grained text mask classification; (b) coarse-grained horizontal text region classification; (c) coarse-grained rotated text region classification; and (d) multi-grain text classification. Each sub-figure is captioned by three text scores for the text from top-to-bottom.

to gain a coarse-grain text score $S_0$, we apply four $3 \times 3$ convolutions, one full connection, and one sigmoid operation to the whole omnidirectional text features.

For the mask reconstruction of arbitrary-shape text, with the omnidirectional pyramid mask proposals $\tilde{Q}$, we firstly calculate all the horizontal rectangle proposals $B$ with Equations 3 and 4.

$$\begin{aligned} QB &= cv2.boxPoints(\tilde{Q}) \\ &= \{X_{1s}, Y_{1s}, X_{2s}, Y_{2s}, X_{3s}, Y_{3s}, X_{4s}, Y_{4s}\} \end{aligned} \tag{3}$$

$$B = \begin{Bmatrix} X_{tl} \\ Y_{tl} \\ X_{br} \\ Y_{br} \end{Bmatrix} = \begin{Bmatrix} min\{X_{1s}, X_{2s}, X_{3s}, X_{4s}\} \\ min\{Y_{1s}, Y_{2s}, Y_{3s}, Y_{4s}\} \\ max\{X_{1s}, X_{2s}, X_{3s}, X_{4s}\} \\ max\{Y_{1s}, Y_{2s}, Y_{3s}, Y_{4s}\} \end{Bmatrix} \tag{4}$$

where, $cv2.boxPoints$ is an OpenCV function which computes the four vertexes of each omnidirectional proposal, $(X_{tl}, Y_{tl}), (X_{br}, Y_{br})$ stand for the top-left and bottom-right coordinates of all proposals separately. Then, with the horizontal proposals $B$, we apply the Skip RoIAlign on the successive feature maps from MASF module to extract text features, since Skip RoIAlign is more efficient during inference while acquiring the comparable performance of Rotated Skip RoIAlign. Finally, we reconstruct the arbitrary-shape text masks with six successively interlaced $\{3 \times 3, 1 \times 1\}$ convolutions and a sigmoid layer. The reconstructed text masks can not only localize the arbitrary-shape texts precisely but also contain the pixel-level text scores, which can be used to calculate a fine-grained text score $S_1$ by Equation 5.

$$S_1 = \frac{1}{\mid R_m \mid} \sum_{(x,y) \in R_m} MS(x, y) \tag{5}$$

where, $R_m$ is the reconstructed arbitrary-shape text mask binarized by threshold 0.5 in Fig. 8(a), $\mid * \mid$ is the L0 norm, $(x, y)$ stands for the pixel coordinate in the text mask, $MS$ describes the predicted text score of each pixel. With $R_m$ of each horizontal proposal, we patch it to a black canvas with size of the original image, and further extract the arbitrary-shape text by OpenCV functions.

To acquire the final multi-grain classification score $S_{mgc}$, we fuse the different grain-size text scores $\{S_0, S_1\}$ by Equation 6.

$$S_{mgc} = \frac{\beta S_0 + (2 - \beta) S_1}{2} \tag{6}$$

where $\beta$ is the weighting coefficient, and it satisfies $0 < \beta < 2$. Obviously, the final multi-grain classification scores $S_{mgc}$ in Fig. 8(d) are more robust.

### D. Optimization

*1) Ground Truths:* Since the features of small texts will vanish in CNN with the forward down-samplings of input images, different stages should comprise different scale texts for our pyramid mask proposal network. We distribute all ground truths $G$ to suitable stages by Equations 7 and 8.

$$min\_sides = min(f(G)[1]) \tag{7}$$

$$T_{lvls} = \begin{cases} 2 & min\_sides > \eta \\ 3 & min\_sides > 2\eta \\ 4 & min\_sides > 3\eta \end{cases} \tag{8}$$

where $f$ is the OpenCV function $cv2.minAreaRect$ and used to calculate the heights and widths of all ground truths, $\eta$ is an empirical value for the specific benchmark, and $T_{lvls}$ denotes the distribution of the ground truths at different stages of Fig. 3. To gain the label of shrunk text region, we conduct the inverse operation of Equation 2 for all quadrilateral ground truths, where different benchmarks have specific values of $s_0$. For instance, $s_0$ is 0.10 in IC15 [49] because texts in IC15 are smaller and closer to each other, larger $s_0$ will reduce the supervision information of masks and smaller $s_0$ is not beneficial for the learning of PLSRSM module. For the arbitrary-shape text in CTW1500 [29], we adopt a label generation strategy similar to that used in TextSnake [50], and $s_0$ is 0.15 because of larger granularity texts in the benchmark.

Text sizes may vary appreciably in scene images. Therefore, if all text pixels contribute equally to the loss function, large text instances will dominative in the loss computation while small ones will be ignored. To tackle this problem, we adopt a pyramid instance-balanced strategy. Specifically, assuming a pyramid layer $l$ in Fig. 3 with $N_l$ text fields, all text fields are treated equally by assigning an equivalent weight $W_{l\_i}$ to each of them. For the i-th text field with area $S_{l\_i}$, each positive pixel within it has a weight $w_{l\_i}$.

$$S_l = \sum_{i}^{N_l} S_{l\_i}, \quad \forall i \in \{1, \ldots, N_l\} \tag{9}$$

$$w_{l\_i} = \frac{W_{l\_i}}{S_{l\_i}}, \quad W_{l\_i} = \frac{S_l}{N_l} \tag{10}$$

All negative pixels are selected by the online hard example mining [51] with three times the number of positive pixels, and the weights are assigned values of 1. For the link prediction branch, a link between two adjacent pixels that belong to the same text is positive, while the link between two adjacent pixels that do not belong to the same text is negative.

In the multi-grain text classification module, all the ground truths are positive proposals for the coarse-grained text classification. Additionally, the scaled omnidirectional proposals $\tilde{Q}$ that overlap with ground truths $G$ over a threshold of 0.5 are positive, while the remaining proposals are negative. For the fine-grained mask reconstruction of arbitrary-shape text, only positive proposals are exploited, pixels within text region of complete polygon rather than shrunk polygon are positive and the remainder are negative.

*2) Loss Functions:* With previously generated labels and weights, the loss function for the omnidirectional pyramid mask proposal subnetwork is composed as Equation 11. The whole loss function of our model is formulated as Equation 12.

$$L_p = \mu_l L_{l\_cls} + \nu_l L_{l\_link},\ \ l \in Stages\ \{2,3,4\} \tag{11}$$

$$Lo = \lambda_p L_p + \lambda_m L_{mask} + \lambda_w L_{w\_cls} \tag{12}$$

where $L_{l\_cls}$ and $L_{l\_link}$ are respectively the losses of pixel classification and pixel-pair link prediction of text proposal generation, $L_{mask}$ and $L_{w\_cls}$ are respectively the losses of the arbitrary-shape mask reconstruction and the whole text classification, and $\{\mu_l, \nu_l, \lambda_p, \lambda_m, \lambda_w\}$ denotes balance factors that harmonize the effects of different losses ($\mu_l = 2, \nu_l = \lambda_p = \lambda_m = \lambda_w = 1$). Besides, the model is trained in a fully end-to-end multi-scale fashion with the above loss $Lo$.

## IV. Experiments

This section compares our OPMP with state-of-the-art methods on several omnidirectional and arbitrary-shape scene text benchmarks.

### A. Benchmarks

*1) ICDAR2015 (IC15) [49]:* IC15 is one of the most popular benchmarks for omnidirectional scene text detection. The images are incidentally captured from streets and shopping malls, and thus challenges of this dataset relate to the omnidirectional, small, and low resolution texts. This dataset contains 1000 training samples and 500 test samples.

*2) MLT [52]:* MLT is a multi-lingual omnidirectional scene text dataset, including 7200 training samples, 1800 validation samples, and 9000 test images. Different annotating styles for different languages and more omnidirectional and perspective distortion texts on various complicated backgrounds make the benchmark challenging.

*3) MSRA-TD500 (TD500) [53]:* TD500 is a text-line based omnidirectional dataset that contains 300 training images and 200 test images captured from indoor and outdoor scenes. Although this dataset has fewer texts per image and most texts are clean, the major challenge is that most texts in the dataset have large variances in orientations.

*4) CTW1500 [29]:* CTW1500 is an arbitrary-shape text detection dataset that contains curved and wavy scene texts. The dataset comprises 1000 training images and 500 test images. Text instances are annotated in a text-line manner with 14-vertex polygons.

*5) Total-Text [54]:* Total-text dataset also aims at detecting the arbitrary-shape texts. It contains 1255 training images and 300 test images. Annotations are given in word level with polygons instead of conventional rectangular bounding boxes.

*6) COCO-Text [55]:* COCO-Text contains 43686 training images, 10000 validation images, and 10000 testing images, respectively. It is very challenging since texts in this dataset are in arbitrary orientations and the annotations are not as accurate as other test datasets in this paper.

### B. Experimental Details

Our method OPMP is built upon the Pytorch framework. In the training phase, the learning rate is initially $10^{-2}$, and reduces to $10^{-3}$ and $10^{-4}$ at the 40th and 60th epochs respectively, the momentum is 0.9, and the overall batch size is 4 on two NVIDIA TITAN X GPUs. The training scales are from 544 to 736 in intervals of 32, and the maximum size of the input image is restricted to 1280 pixels. We pretrain our model on the MLT training samples. The pretrained model is then fine-tuned on each specific dataset. The batch size of the omnidirectional proposals fed to the arbitrary-shape text detection stage is 128. Apart from the original ground truth proposals and the proposals generated by the omnidirectional pyramid mask proposal subnetwork, parts of the positive omnidirectional proposals are randomly sampled around the text regions according to the ground truths, and parts of the negative omnidirectional proposals are randomly generated by applying the criterion of the OHEM [51] to the difficult background regions. For testing, we filter the results by using the shorter side length, area, and confidence, which have values of 10, 150, and 0.85 respectively.

### C. Quantitative Evaluation

Existing state-of-the-art methods have various experiment settings, for example, exploiting different network backbones (e.g. VGG16 [56], ResNet-50 [40], DenseNet [57]), multi-scale training and testing, data augmentation [58], pretraining with different extra datasets, which is hard to conduct a full fair comparison. Therefore, we compare our method with other state-of-the-art methods as fairly as possible during the quantitative evaluation.

*1) Evaluation on IC15:* The whole test process is conducted for the single-scale $768 \times 1280$ case. In Table I, compared with the Mask R-CNN algorithm [14], which suffers from the stack-omnidirectional text dilemma due to the suppression of redundant horizontal proposals with NMS [16], our OPMP directly generates a suitable number of omnidirectional pyramid mask proposals without the NMS procedure, obviously solving the problem described above. It is well known that IC15 is more challenging in terms of the separation of adjacent texts, image quality degradation, and complex background. Table I shows that our OPMP is better than state-of-the-art methods in terms

TABLE I
COMPARISON WITH CLASSICAL METHODS ON ICDAR2015. R: RECALL, P: PRECISION, AND F: F-SCORE. "⁎": THE NETWORK BACKBONE IS RESNET-50. THE RUNTIME FOR ALL METHODS OTHER THAN OUR'S ARE OBTAINED FROM THE REPORTED PAPERS ACCORDINGLY

| Methods | R | P | F | FPS | TIoU-Hmean[11] |
|---|---|---|---|---|---|
| DMPNet [10] | 68.2% | 73.2% | 70.6% | | 53.2% |
| WordSup [59] | 77.0% | 79.3% | 78.2% | 2 | 55.3% |
| DDR [31] | 80% | 82% | 81% | 1.1 | - |
| EAST [12] | 78.3% | 83.2% | 80.7% | **13.2** | 60.1% |
| RRPN [17] | 77% | 84% | 80% | 3 | - |
| Mask R-CNN⁎ [14] | 81.5% | 83.8% | 82.6% | 5 | 59.3% |
| RRD [60] | 79% | 85.6% | 82.2% | 6.5 | 58.5% |
| TextSnake [50] | 80.4% | 84.9% | 82.6% | 1.1 | - |
| TextBox++ [61] | 76.7% | 87.2% | 81.7% | 11.6 | 60.3% |
| TextField [62] | 80.5% | 84.3% | 82.4% | 5.2 | - |
| PixelLink [7] | 82.0% | 85.5% | 83.7% | 3 | 60.5% |
| IncepText⁎ [63] | 80.6% | 90.5% | 85.3% | 3.7 | - |
| AF-RPN [39] | 83% | 90% | 86.3% | 5 | - |
| SPCNET ⁎ [64] | **85.8%** | 88.7% | 87.2% | 2 | - |
| SBD⁎ [65] | 83.8% | 89.4% | 86.5% | 3.2 | - |
| MSR⁎ [66] | 78.4% | 86.6% | 82.3% | 4.3 | - |
| ICG [67] | 80.3% | 83.7% | 82.0% | 7.1 | - |
| LOMO⁎ [68] | 83.5% | **91.3%** | 87.2% | | - |
| our OPMP⁎ | 85.5% | 89.1% | **87.3%** | 1.4 | 63.4% |

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON MSRA-TD500

| Methods | R | P | F | FPS |
|---|---|---|---|---|
| RRPN [17] | 67.0% | 71.8% | 69.3% | 3 |
| DDR [31] | 70% | 77% | 74% | 1.1 |
| EAST [12] | 67.4% | 87.3% | 76.1% | **13.2** |
| RRD [60] | 73% | 87% | 79% | 6.5 |
| BSAB [70] | 77.4% | 83.0% | 80.1% | |
| PixelLink [7] | 73% | 81.1% | 76.8% | 3 |
| Mask R-CNN [14] | 74.9% | 82.8% | 78.7% | 5 |
| TextSnake [50] | 73.9% | 83.2% | 78.3% | 1.1 |
| IncepText [63] | 79.0% | 87.5% | 83.0% | 3.7 |
| TextField [62] | 75.9% | 87.4% | 81.3% | |
| SBD [65] | 80.5% | **89.6%** | **84.8%** | 3.2 |
| MSR [66] | 76.7% | 87.4% | 81.7% | |
| our OPMP | **83.4%** | 86.0% | 84.7% | 1.6 |

of the F-score, revealing that our method is more robust in these cases. In particular, compared with PixelLink [7], which suffers from the under-segmentation of very close text words, our OPMP solves the problem with a $3.6\%$ performance enhancement in the F-score.

*2) Evaluation on TD500:* TD500 is labeled by text lines and the major challenge is therefore that the texts have large variances in orientations and aspect-ratios, resulting in many state-of-the-art methods suffering from the over-segmentation problem shown in Fig. 1(c). Similar to some previous state-of-the-art methods, such as RRPN [17], PixelLink [7], etc., we use the HUST-TR400 [69] dataset to expand the training samples owing to the limited number of samples in TD500. The test resolution is restricted to $768 \times 1280$ pixels. Table II shows that our OPMP well outperforms other instance segmentation methods [7] [62], strongly indicating the greater effectiveness of our method in solving the problem of the over-segmentation of long text lines.

*3) Evaluation on Arbitrary-Shape Texts:* As CTW1500 and Total-text are labeled by arbitrary-shape text lines and words, respectively, it is challenging in terms of the large variations in orientations and shapes, raising the stack-omnidirectional text dilemma for the methods of Mask R-CNN [14], CTD [29], and SLPR [71]. For instance, some texts enclosed by long text lines

TABLE III
COMPARISON BETWEEN STATE-OF-THE-ART METHODS AND OUR METHOD OPMP ON ARBITRARY-SHAPE SCENE TEXTS OF CTW1500 AND TOTAL-TEXT DATASETS. MS: MULTI-SCALE TEST. "★" AND "⁎": THE NETWORK BACKBONES ARE VGG16 AND RESNET-50. SYN [72] AND MLT ARE THE PRETRAINING DATASETS

| Methods | CTW1500, $\beta = 0.6, s_0 = 0.15$ | | | |
|---|---|---|---|---|
| | R | P | F | FPS |
| CTD [29] | 65.2% | 74.3% | 69.5% | **15.2** |
| CTD + TLOC [29] | 69.8% | 74.3% | 73.4% | 13.3 |
| SLPR [71] | 70.1% | 80.1% | 74.8% | |
| Mask R-CNN [14] | 75.7% | 81.5% | 78.5% | 5 |
| TextSnake [50] | 85.3% | 67.9% | 75.6% | 1.1 |
| CSE [73] | 76.0% | 81.1% | 78.4% | 2.6 |
| MSR [66] | 78.3% | 85.0% | 81.5% | |
| ICG [67] | 79.8% | 82.8% | 81.3% | |
| LOMO [68] | 69.6% | **89.2%** | 78.4% | 4.4 |
| LOMO MS [68] | 76.5% | 85.7% | 80.8% | |
| TextField [62] | 79.8% | 83.0% | 81.4% | |
| our OPMP | 80.8% | 85.1% | **82.9%** | 1.4 |
| Methods | Total-text, $\beta = 0.6, s_0 = 0.15$ | | | |
| | R | P | F | FPS |
| CTD + TLOC [29] | 71.0% | 74.0% | 73.0% | **13.3** |
| Mask R-CNN⁎ [14] + MLT | 77.4% | 82.1% | 79.7% | 5 |
| TextSnake★ [50] + Syn | 74.5% | 82.7% | 78.4% | 1.1 |
| TextField★ [62] + Syn | 79.9% | 81.2% | 80.6% | |
| SPCNET ⁎ [64] + Syn | 82.8% | 83.0% | 82.9% | 2 |
| MSR⁎ [66] + Syn | 74.8% | 83.8% | 79.0% | |
| ICG [67] + Syn | 80.9% | 82.1% | 81.5% | |
| LOMO⁎ MS [68] + Syn | 79.3% | 87.6% | 83.3% | |
| our OPMP⁎ + MLT | **82.9%** | **88.5%** | **85.6%** | 1.4 |
| Methods | Total-text, $\beta = 0.6, s_0 = 0.15$ | | | |
| | R | P | F | FPS |
| OPMP★ + Syn | 80.3% | 85.2% | 82.7% | 3.7 |
| OPMP⁎ + Syn | 82.7% | 87.6% | 85.1% | 1.4 |
| OPMP⁎ + MLT | **82.9%** | **88.5%** | **85.6%** | 1.4 |

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS ON MLT BENCHMARK. MS: MULTI-SCALE TEST

| Methods | Recall | Precision | F-score | FPS |
|---|---|---|---|---|
| linkage-ER-Flow [52] | 25.6% | 44.5% | 32.5% | |
| TH-DL [52] | 34.8% | 67.8% | 46.0% | |
| SARI_FDU_RRPN_v1 [52] [17] | 55.5% | 71.2% | 62.4% | |
| Sensetime OCR [52] | 69.4% | 56.9% | 62.6% | |
| SCUT_DLVClab [52] | 54.5% | 80.3% | 65.0% | |
| DDR [52] [31] | 57.9% | 76.7% | 66.0% | |
| CLRS [74] | 55.6% | **83.8%** | 66.8% | |
| AF-RPN [39] | 66% | 75% | 70% | |
| BSAB [70] | 62.1% | 77.7% | 69.0% | |
| FOTS [75] | 62.3% | 82.9% | 70.8% | |
| Mask R-CNN [14] | 63.7% | 81% | 71.3% | |
| SPCNET [64] | 66.9% | 73.4% | 70.0% | |
| SPCNET MS [64] | 68.6% | 80.6% | 74.1% | |
| SBD [65] | 70.1% | 83.6% | **76.3%** | 3.2 |
| our OPMP | **70.5%** | 82.9% | 76.2% | 0.9 |

are easily suppressed by the NMS algorithm. Additionally, the large aspect-ratios of arbitrary-shape texts in CTW1500 also make numerous state-of-the-art methods suffer from the over-segmentation problem easily because of limited receptive fields. During the test phase, the minimum and maximum sides of input are respectively 768 and 1280 in length. Table III clearly shows that our OPMP addresses the above problems more robustly, and the speed of our model is comparable to the speeds of other state-of-the-art methods.

*4) Evaluation on MLT:* During the test procedure, the minimum and maximum sizes of the dataset are 960 and 1600 respectively, which is different from the setting of SBD [65] (minimum

TABLE V
THE EFFECT OF DIFFERENT COMPONENTS OF OUR METHOD. PLSRSM: PYRAMID LENGTHWISE AND SIDEWISE RESIDUAL SEQUENCE MODELING. MASF: MULTIPLE ARBITRARY-SHAPE FITTING. MGTC: MULTI-GRAIN TEXT CLASSIFICATION. "♦": WITHOUT REVERSE OPERATION IN LSTM. LINKS: LINKS FOR 8-NEIGHBOUR PIXEL PAIRS. SP: GROUND-TRUTH WITH SHRUNK POLYGON

| Methods | ICDAR2015, $\beta = 1.3, s_0 = 0.10$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | R | P | F | ∇R | ∇P | ∇F | FPS |
| our OPMP | 85.5% | 89.1% | **87.3%** | | | | 1.4 |
| our OPMP♦ | 85.2% | 88.1% | 86.6% | 0.3% ↓ | 1.0% ↓ | 0.7% ↓ | 1.46 |
| OPMP - PLSRSM | 84.7% | 86.6% | 85.6% | 0.8% ↓ | **2.5%** ↓ | 1.7% ↓ | 1.5 |
| OPMP - MASF | 84.1% | 86.8% | 85.4% | 1.4% ↓ | 2.3% ↓ | **1.9%** ↓ | 1.5 |
| OPMP - MGTC | 83.9% | **89.3%** | 86.5% | **1.6%** ↓ | 0.2% ↑ | 0.8% ↓ | 1.6 |
| Methods | CTW1500, $\beta = 0.6, s_0 = 0.15$ | | | | | | |
| | R | P | F | ∇R | ∇P | ∇F | FPS |
| our OPMP | **80.8%** | **85.1%** | **82.9%** | | | | 1.4 |
| OPMP - PLSRSM | 80.5% | 83.3% | 81.9% | 0.3% ↓ | **1.8%** ↓ | 1.0% ↓ | 1.5 |
| OPMP - MASF | 79.5% | 83.4% | 81.4% | 1.3% ↓ | 1.7% ↓ | **1.5%** ↓ | 1.5 |
| OPMP - MGTC | 79.3% | 85.0% | 82.1% | **1.5%** ↓ | 0.1% ↓ | 0.8% ↓ | 1.6 |
| Methods | ICDAR2015, $\beta = 1.3, s_0 = 0.10$ | | | | | | |
| | R | P | F | ∇R | ∇P | ∇F | FPS |
| our OPMP | 85.5% | 89.1% | **87.3%** | | | | 1.4 |
| OPMP - Links | 84.8% | 87.3% | 86.0% | **0.7%** ↓ | **1.8%** ↓ | **1.3%** ↓ | 1.5 |
| OPMP - SP | 85.1% | 88.4% | 86.7% | 0.4% ↓ | 0.7% ↓ | 0.6% ↓ | 1.35 |

TABLE VI
THE EFFECT OF OUR MASF MODULE. OPMP†: REPLACING THE MULTI-IMPROVED DEFORMABLE CONVOLUTION BY THE TRADITIONAL DEFORMABLE CONVOLUTION [46] IN OUR MASF MODULE

| Methods | CTW1500, $\beta = 0.6$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | R | P | F | ∇R | ∇P | ∇F | FPS |
| our OPMP | 80.8% | 85.1% | 82.9% | | | | 1.4 |
| our OPMP† | 79.9% | 84.3% | 82.0% | 0.9% ↓ | 0.8% ↓ | 0.9% ↓ | 1.45 |

TABLE VII
THE EFFECT OF THE PROPOSAL NUMBER ON THE RECALL RATE. SYMBOL *: ONLY INCLUDING THE PROPOSAL GENERATION STAGE. R: RECALL. AR: AVERAGE RECALL AT MULTIPLE IOU THRESHOLDS BETWEEN 0.50 AND 0.95 WITH AN INTERVAL OF 0.05

| Methods | Proposals number | COCO-Text validation | | |
|---|---|---|---|---|
| | | R (IoU=0.5) | R (IoU=0.75) | AR |
| Mask R-CNN* | 100 | 78.7% | 37.4% | 38.8% |
| AF-RPN* [39] | 100 | 81.8% | 41.3% | 43.6% |
| our OPMP* | $< 100$ | 82.6% | 43.2% | 45.1% |

size is 1200). Results are given in Table IV. Our method outperforms many state-of-the-art methods by a large margin in the single-scale case, suggesting that our method has a strong generalization ability on a huge multiple language benchmark.

*5) Ablation Studies:* Table V presents the results of comparison experiments conducted on two benchmarks that verify the effectiveness of different components of our method. The symbol "-" stands for the removal of a certain component from our best model, "↑" and "↓" describe the percentage increase and decline respectively, and $(\nabla R, \nabla P, \nabla F)$ denotes the absolute percentage differences of the recall (R), precision (P), and F-score (F) metrics. Removing the multi-grain text classification module means calculating each final text score via only using the reconstructed text mask. Table V clearly shows the followings. (1) Via the clear improvement in precision, our PLSRSM module solves the problems of the under-segmentation of very close text words and the over-segmentation of long text lines well as in Fig. 1, because the module strengthens or weakens the relationship among broad context features through its inherent strong control of space information flow. And Fig. 4 also testifies that, since the difference values between foreground and background feature values of very close text words are enlarged, by contrary, the background features between different words of each long text line are weakened to strengthen the relationship. Besides, we further verify the effect of "shrunk polygon" ground-truths and the links for 8-neighbour pixel pairs of PLSRSM module respectively, which proves the main impact of the proposed PLSRSM module. And the reasons are that the blanks generated by "shrunk polygon" between adjacent words will vanish along with further down-sampling in $\{Stage\ 3, Stage4\}$ of Fig. 3, while the broken links between adjacent words still exist because the two pixels connected by one link belong to two close text instances, separately. Additionally, with the several advantages described in subsection III-B, the module hardly consumes much extra time in the light of FPS metric in Table V; (2) Our multiple arbitrary-shape fitting module enhances both the recall and precision because of its powerful ability to model arbitrary-shape text information. (3) With different optimal $\beta$ values for different benchmarks, the proposed multi-grain text classification module can further raise the recall of our OPMP for a high confidence threshold, which shows that the module takes full advantage of different grain-size classifications to realize robust text scores for all omnidirectional proposals. Additionally, in Table VI, because we leverage the large receptive field and irregular offset prediction rather than the regular offset prediction, the improved deformable convolution in our MASF module allows for the much better detection of arbitrary-shape texts.

Table VII presents a series of ablation experiments conducted on COCO-Text [55] to evaluate the region proposal quality of anchor-based and anchor-free methods. As the original RPN in Mask R-CNN [14] can not output quadrilateral proposals without an omnidirectional anchor design, we evaluate the axis-aligned rectangular proposal quality for a fair comparison. We compute the recall rates $R_{(IoU=0.5)}$ and $R_{(IoU=0.75)}$ at single intersection over union (IoU) thresholds of 0.5 and 0.75 respectively. Additionally, for the number of text proposals, because our OPMP generates only a suitable number of proposals ($< 100$) rather than the massive number of redundant proposals in other methods, we select the first 100 proposals from other state-of-the-art methods for reasonable comparison. Obviously, in Table VII, although our OPMP generates fewer proposals than others, it has an evident recall gain at multiple IoU thresholds, which strongly indicates the effectiveness of our generated omnidirectional pyramid mask proposals and the proposed novel PLSRSM module.

In summary, all results of experiments in Tables V, VI and VII show the remarkable effectiveness and efficiency of each innovative component in our OPMP.

*6) The Influence of Different Network Backbones and Pretraining Datasets:* Since numerous state-of-the-art methods conduct experiments with different network backbones and pretraining datasets, we also verify the effectiveness of our OPMP with similar experiment settings in Table III for a relative fairness comparison. The results show that our method can outperform previous methods by a large margin. Besides, the influence


(a) ICDAR2015 (b) MLT (c) MSRA-TD500 (d) CTW1500

Fig. 9. Qualitative evaluation results. The four columns present detection results for the ICDAR2015, MLT, MSRA-TD500, and CTW1500 benchmarks. Each detected text instance is visualized by its confidence, arbitrary-shape mask proposal, the contour (red line) or external quadrilateral (blue line) of the reconstructed mask, ground-truth (yellow), ignorable ground-truth (cyan), and the green bounding-boxes are auxiliary to display the confidences.

of different pretraining datasets is not remarkably, and there are several reasons leading it. Firstly, although the MLT [52] dataset has 9000 training and validation samples captured from the wild, its sample number is far fewer than the sample number (800000) of Syn [72]. Secondly, numerous samples with other languages rather than English in MLT [52] are not suitable for pretraining models for specific text detection tasks. However, even though with the same experiment settings, our method still outperforms other state-of-the-art methods by a large margin, which demonstrates the robustness of our method.

### D. Qualitative Evaluation

Fig. 9 presents a selection of qualitative experimental results for the four public benchmarks, with the four columns giving the detection results for ICDAR2015, MLT, MSRA-TD500, and CTW1500 benchmarks. Meanwhile, each detection result is visualized by its confidence, omnidirectional proposal mask, and the final detected quadrilateral or contour of the reconstructed mask, which is generated in the Arbitrary-shape Text Detection stage. Our approach performs well for the variously challenging scene text images. Specifically, the first column for ICDAR2015 shows that our method well separates texts in the very adjacent and small text instances, especially for the omnidirectional proposal masks. This demonstrates the effectiveness of our method in solving the under-segmentation problem for very adjacent text words. The second column testifies that our detector detects multiple language scene texts finely. The third and fourth columns show that our OPMP detects very long and arbitrary-shape text lines robustly, further certifying the effectiveness of our method in solving the over-segmentation problem for long text-lines with the large aspect ratios.

### E. Weaknesses

As demonstrated in the preceding experiments, our OPMP works well in most cases of detecting arbitrary-shape texts. It fails for some difficult images, such as images having object occlusion, and text-like objects, which are common challenges for other state-of-the-art methods [12], [32], [61], [62]. Examples of failure are displayed in Fig. 10.


Fig. 10. Examples of failure. Green contours: correct detections; Red contours: missing detection; Blue contours: false detections; Yellow contours: missing ground truths in the test set of the given benchmark.

## V. Conclusion

We have proposed an omnidirectional pyramid mask proposal text detector termed OPMP, which has several advantages. (1) OPMP solves the problem of the stack-omnidirectional text dilemma faced by some traditional methods through a suitable number of omnidirectional pyramid mask proposals. (2) OPMP addresses the problems of the under-segmentation of very close text words and the over-segmentation of long text lines with an innovative PLSRSM module. (3) OPMP accurately and quickly fits the shapes of arbitrary-shape scene texts with a novel multiple arbitrary-shape fitting module. (4) A multi-grain text classification module is proposed to classify each arbitrary-shape text with a robust text score. Comprehensive ablation studies and experiments on various public text detection benchmarks all demonstrate the effectiveness of OPMP. In the future, we will try and extend the OPMP to an end-to-end scene text spotting framework.

## References

[1] 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.

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

[3] Y. Tang and X. Wu, "Scene text detection using superpixel-based stroke feature transform and deep learning based region classification," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2276–2288, Sep. 2018.

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

[5] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1480–1500, Jul. 2015.

[6] S. Zhang, Y. Liu, L. Jin, and C. Luo, "Feature enhancement network: A refined scene text detector," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 2612–2619.

[7] 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.

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

[9] H. Li, P. Wang, and C. Shen, "Towards end-to-end text spotting with convolutional recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5238–5246.

[10] Y. Liu and L. Jin, "Deep matching prior network: Toward tighter multi-oriented text detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1962–1969.

[11] Y. Liu, L. Jin, Z. Xie, C. Luo, S. Zhang, and L. Xie, "Tightness-aware evaluation protocol for scene text detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 9612–9620.

[12] X. Zhou *et al.*, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017.

[13] 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.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2961–2969.

[15] Z. Zhong, L. Jin, and S. Huang, "DeepText: A new approach for text proposal generation and text detection in natural images," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 1208–1212.

[16] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. IEEE Pattern Recognit., 18th Int. Conf.*, vol. 3, 2006, pp. 850–855.

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

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

[19] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2010, pp. 2963–2970.

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

[21] W. Huang, Z. Lin, J. Yang, and J. Wang, "Text localization in natural images using stroke feature transform and text covariance descriptors," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 1241–1248.

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

[23] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 3538–3545.

[24] K. Wang and S. Belongie, "Word spotting in the wild," in *Proc. Eur. Conf. Comput. Vision*, 2010, pp. 591–604.

[25] 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.

[26] L. Gomez and D. Karatzas, "Multi-script text extraction from natural scenes," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, 2013, pp. 467–471.

[27] Y. Li, W. Jia, C. Shen, and A. van den Hengel, "Characterness: An indicator of text in the wild," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1666–1677, Apr. 2014.

[28] H. Li, P. Wang, and C. Shen, "Toward end-to-end car license plate detection and recognition with deep neural networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 1126–1136, Mar. 2018.

[29] Y. Liu, L. Jin, S. Zhang, C. Luo, and S. Zhang, "Curved scene text detection via transverse and longitudinal sequence connection," *Pattern Recognit.*, vol. 90, pp. 337–345, 2019.

[30] C. Yan *et al.*, "A fast uyghur text detector for complex background images," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3389–3398, Dec. 2018.

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

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

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

[34] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2550–2558.
[35] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
[36] C. Yao *et al.*, "Scene text detection via holistic, multi-channel prediction," 2016, *arXiv:1606.09002*.
[37] D. He *et al.*, "Multi-scale FCN with cascaded instance aware segmentation for arbitrary oriented word spotting in the wild," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 3519–3528.
[38] Y. Wu and P. Natarajan, "Self-organized text detection with minimal post-processing via border learning," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5000–5009.
[39] Z. Zhong, L. Sun, and Q. Huo, "An anchor-free region proposal network for faster R-CNN based text detection approaches," *Int. J. Document Anal. Recognit.*, vol. 22, no. 3, pp. 315–327, 2019.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.
[41] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2117–2125.
[42] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 3–19.
[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[45] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "AON: Towards arbitrarily-oriented text recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 5571–5579.
[46] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 764–773.
[47] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[48] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2874–2883.
[49] D. Karatzas *et al.*, "ICDAR 2015 competition on robust reading," in *Proc. IEEE Document Anal. Recognit., 13th Int. Conf.*, 2015, pp. 1156–1160.
[50] 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. Vision*, 2018, pp. 20–36.
[51] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 761–769.
[52] N. Nayef *et al.*, "ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification-RRC-MLT," in *Proc. IEEE Document Anal. Recognit., 14th IAPR Int. Conf.*, 2017, vol. 1, pp. 1454–1459.
[53] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 1083–1090.
[54] C. K. Ch'ng and C. S. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in *Proc. IEEE 14th IAPR Int. Conf. Document Anal. Recognit.*, 2017, vol. 1, pp. 935–942.
[55] 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*.
[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[57] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 4700–4708.
[58] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 21–37.
[59] H. Hu *et al.*, "WordSup: Exploiting word annotations for character based text detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017.
[60] 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. Vision Pattern Recognit.*, 2018, pp. 5909–5918.
[61] 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.
[62] Y. Xu *et al.*, "TextField: Learning a deep direction field for irregular scene text detection," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5566–5579, Nov. 2019.
[63] Q. Yang *et al.*, "IncepText: A new inception-text module with deformable psroi pooling for multi-oriented scene text detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 1071–1077.
[64] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao, and G. Li, "Scene text detection with supervised pyramid context network," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 9038–9045.
[65] Y. Liu *et al.*, "Omnidirectional scene text detection with sequential-free box discretization," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 3052–3058.
[66] C. Xue, S. Lu, and W. Zhang, "Msr: Multi-scale shape regression for scene text detection," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, AAAI Press, 2019, pp. 989–995.
[67] J. Tang *et al.*, "Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping," *Pattern Recognit.*, vol. 96, no. 10, pp. 6954–6966, 2019.
[68] C. Zhang *et al.*, "Look more than once: An accurate detector for text of arbitrary shapes," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 10 552–10 561.
[69] C. Yao, X. Bai, and W. Liu, "A unified framework for multioriented text detection and recognition," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4737–4749, Nov. 2014.
[70] C. Xue, S. Lu, and F. Zhan, "Accurate scene text detection through border semantics awareness and bootstrapping," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 355–372.
[71] Y. Zhu and J. Du, "Sliding line point regression for shape robust scene text detection," in *Proc. 24th Int. Conf. Pattern Recognit.*, 2018, pp. 3735–3740.
[72] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2315–2324.
[73] Z. Liu *et al.*, "Towards robust curve text detection with conditional spatial expansion," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 7269–7278.
[74] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7553–7563.
[75] X. Liu *et al.*, "FOTS: Fast oriented text spotting with a unified network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 5676–5685.

**Sheng Zhang** received the B.S. degree from Hohai University, Nanjing, China, in 2014. He is currently working toward the Ph.D. degree with DLVC Laboratory, South China University of Technology, Guangzhou, China. His major interests include research on object tracking, deep learning, and image processing algorithms.

**Yuliang Liu** received the B.S. degree from the South China University of Technology, Guangzhou, China, in 2016, where he is currently working toward the Ph.D. degree and a Visitor Ph.D. student with the University of Adelaide, Adelaide, SA, Australia. His research interests include deep learning, object detection, and scene text reading.

**Lianwen Jin** (Member, IEEE) received the B.S. degree from the University of Science and Technology of China, Anhui, China, and the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 1991 and 1996, respectively. He is a Professor with the College of Electronic and Information Engineering, South China University of Technology. He has authored more than 100 scientific papers. His research interests include handwriting analysis and recognition, image processing, machine learning, and intelligent systems. He was the recipient of the New Century Excellent Talent Program of MOE Award and the Guangdong Pearl River Distinguished Professor Award, and is a member of the IEEE Computational Intelligence Society, IEEE Signal Processing Society, and IEEE Computer Society.

**Zhongrong Wei** received the bachelor of engineering degree from the Beijing Institute of Technology, Beijing, China, in 2018. He is currently working toward the master's degree with the DLVC laboratory, South China University of Technology, Guangzhou, China. His major interests include research on deep learning, image processing, and computer vision.

**Chunhua Shen** is currently a Professor with The University of Adelaide, Adelaide, SA, Australia.