Zoom Text Detector

Chuang Yang, Mulin Chen, Yuan Yuan, *Senior Member, IEEE,* and Qi Wang, *Senior Member, IEEE*

***Abstract*—To pursue comprehensive performance, recent text detectors improve detection speed at the expense of accuracy. They adopt shrink-mask based text representation strategies, which leads to a high dependency of detection accuracy on shrink-masks. Unfortunately, three disadvantages cause unreliable shrink-masks. Specifically, these methods try to strengthen the discrimination of shrink-masks from the background by semantic information. However, the feature defocusing phenomenon that coarse layers are optimized by fine-grained objectives limits the extraction of semantic features. Meanwhile, since both shrink-masks and the margins belong to texts, the detail loss phenomenon that the margins are ignored hinders the distinguishment of shrink-masks from the margins, which causes ambiguous shrink-mask edges. Moreover, false-positive samples enjoy similar visual features with shrink-masks. They aggravate the decline of shrink-masks recognition. To avoid the above problems, we propose a Zoom Text Detector (ZTD) inspired by the zoom process of the camera. Specifically, Zoom Out Module (ZOM) is introduced to provide coarse-grained optimization objectives for coarse layers to avoid feature defocusing. Meanwhile, Zoom In Module (ZIM) is presented to enhance the margins recognition to prevent detail loss. Furthermore, Sequential-Visual Discriminator (SVD) is designed to suppress false-positive samples by sequential and visual features. Experiments verify the superior comprehensive performance of ZTD.**

***Index Terms*—Text detection, zoom strategy, feature defocusing, detail loss, false-positive samples.**

I. Introduction

TEXT detection, the key to retrieving texts, has become an attractive topic and involves various applications (such as multilingual translation systems and unmanned systems). In the past decade, since deep learning technologies [1]–[3] have shown impressive performance in computer vision and artificial intelligence, many deep learning-based algorithms are proposed for text detection [4], which can be categorized into two classes roughly: accuracy prior methods [5], [7] and comprehensive performance prior methods [8]–[10].

The former represents text instances by multiple local units or rebuilds text contours by a series of geometry operations. These methods usually enjoy high detection accuracy. However, the complicated frameworks lead to expensive memory overhead, deep dependency for high-performance computing units, and slow inference. Furthermore, related works [11]–[13] show the weak gains for model accuracy with the increase of model complexity. The latter aims to accelerate the inference process with lightweight frameworks to make it possible to deploy text detection techniques into mobile terminals. These works [9], [10] model the whole text instances directly through shrink-mask based text representation strategies. They only need to conduct prediction tasks on one feature map that is fused by multi-level feature maps and rebuild text contours by simple post-processing, which simplifies the frameworks and improves the detection speed effectively. However, three disadvantages exist in these methods, which limit the improvement of detection accuracy.

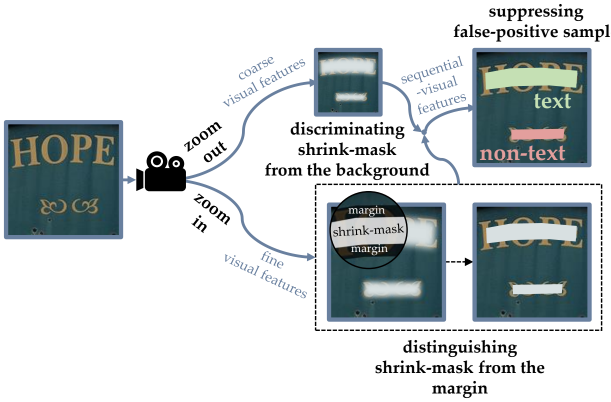


Fig. 1. Motivation of the proposed method. We aim to avoid feature defocusing and detail loss by simulating the zoom process of the camera. Meanwhile, we suppress false-positive samples according to the combination of sequential and visual features.

The first one is the phenomenon of feature defocusing. Shrink-mask based text representation strategies lead to the detection accuracy being highly dependent on shrink-masks. To pursue reliable shrink-masks, current algorithms merge coarse layers into fine layers following the idea of [1]. They try to enhance the discrimination of shrink-masks from the background by the semantic information from coarse layers. However, the layers are only supervised by the fine-grained optimization objectives in the training stage, which limits the extraction of semantic features. The second one is the phenomenon of detail loss. It is found in the Contour Extension Process in Fig. 3, both shrink-masks and the corresponding margins belong to texts, which makes it hard to determine clear borders between shrink-masks and the margins. Existing methods ignore the margins recognition. It accelerates the decline of the model's ability to distinguish shrink-masks from the margins, which results in ambiguous shrink-mask edges and leads to many adverse effects (such as text adhesion, miss detection, and the introduction of noise information). Moreover, false-positive samples enjoy highly similar visual features (such as color, texture, geometry, etc.) with shrink-masks. However, previous detectors suppress them according to visual features only, which aggravates the decline of model accuracy.

Considering the limitations above, how to overcome those

Chuang Yang is with the School of Computer Science, and with the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China. Mulin Chen, Yuan Yuan, and Qi Wang are with the School of Artificial Intelligence, OPtics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China. E-mail: cyang113@mail.nwpu.edu.cn, chenmulin@mail.nwpu.edu.cn, y.yuan.ieee@gmail.com, crabwq@gmail.com. Qi Wang is the corresponding author.

problems is still under explored. The photographers capture global information of scenes by zooming out the camera, which helps to analyze the semantic relationships between different objects. Meanwhile, they zoom-in the camera to focus on local regions, which supports observing object details. In this paper, inspired by the zoom process of the camera, we propose Zoom Text Detector (ZTD). It makes full use of the advantages of coarse and fine features to enhance the reliability of shrink-masks. Specifically, as shown in Fig. 1, to help discriminate shrink-masks from the background, a Zoom Out Module (ZOM) simulates the zoom-out process of the camera to focus on coarse features. It provides coarse-grained optimization objectives for coarse layers to facilitate the extraction of coarse features with strong semantic information. Meanwhile, to strengthen the distinguishment of shrink-masks from the corresponding margins, a Zoom In Module (ZIM) simulates the zoom-in process of the camera to utilize fine features to to enhance ZTD's ability to recognize the margins. Moreover, considering shrink-masks are equipped with rich sequential features, a Sequential-Visual Discriminator (SVD) is designed to combine sequential and visual features to help suppress false-positive samples. The main contributions of this paper are as follows:

1) Zoom strategy-based Zoom Out Module (ZOM) and Zoom In Module (ZIM) are proposed, which maximize the advantages of coarse and fine features to avoid feature defocusing and detail loss. The former helps to discriminate shrink-masks from the background. The latter strengthens the distinguishment of shrink-masks from the margins.
2) A Sequential-Visual Discriminator (SVD) is designed to encourage ZTD to learn the combination of sequential and visual features, which helps to suppress false-positive samples from temporal and spatial domains. Particularly, it brings no computational cost to the inference process and can be integrated into other detectors seamlessly.
3) An efficient text detection framework combined with a lightweight CNN model and simple post-processing is constructed. It achieves the detection accuracy comparable with accuracy prior methods and runs faster than comprehensive performance prior algorithms, which provides sufficient support for practical applications.

The rest of the paper is organized as follows. Section II introduces the related works on text detection. Section III describes the structure of ZTD. The experimental results are discussed in Section IV. Section V concludes the paper.

II. Related Work

In recent years, deep learning-based methods have achieved dominant performance on text detection, which can be classified into accuracy prior methods and comprehensive performance prior methods roughly.

A. Accuracy Prior Methods

Two-stage detection methods, such as Faster-RCNN [14], have brought great inspiration to text detection [15]–[17] early. To speed up the text inference process, more one-stage methods [18], [19] based on SSD [20] were proposed. For extracting text features with strong expression, Liao *et al.* [21] proposed 1×5 convolution kernel to fit text geometries, and He *et al.* [22] introduced attention mechanism into the text detection model. To detect multi-oriented texts, several extra works [16], [23], [24] predicted angles of multi-oriented texts. Others [25], [26] regressed the offsets between bounding boxes and anchors. Since anchor mechanism increases the model complexity largely, anchor-free text detector [27] based on [3], [28], [29] has been proposed, which obtained multiple contour points through regression strategy directly. Recently, researchers focus on the representation of irregular-shaped text contours. Some works [12], [30], [31] detected multiple character-level bounding boxes by regression strategy and linked the boxes to obtain the final text contours. Different from them, other works [32], [33] predicted character-level boxes based on segmentation technology. Furthermore, the authors [34], [35] and [27], [36], [37] modeled text contours as a series of dense contour key points. They only needed to obtain the points through segmentation and regression technologies, respectively. In addition, latest works [11] introduced mathematical model to fit text instances. Specifically Zhu *et al.* [11] modeled text instances in the Fourier domain and expressed arbitrary-shaped text contours as compact signatures. Except for the above algorithms, Tian *et al.* [38] and Xu *et al.* [39] segmented the whole text regions directly and separated adjacent texts by direction maps. Though these methods achieve high detection accuracy for arbitrary-shaped text detection, the complicated framework hinders the feasibility in practical applications.

B. Comprehensive Performance Prior Methods

To meet the high requirements of the practical application scenarios for the comprehensive performance of algorithms, some approaches [8]–[10] improved detection speed at the expense of accuracy by simplifying the framework. Inspired by [28], Zhou *et al.* [8] presented a one-stage text detection method, which optimized the inference process largely after abandoning anchor mechanism. Moreover, the authors augmented positive samples in the training stage by introducing the shrink-mask. Recently, to further simplify the framework while handling irregular-shaped texts, most works are dedicated to researching an efficient text representation model. Wang *et al.* [9], [40] proposed a faster region extension strategy based on their previous work [41], which can effectively fit arbitrary-shaped text instances and overcome the text adhesion problem. Although the methods enjoy high detection speed, pixel-wise extension based post-processing is relatively time-consuming. Considering this problem, Liao *et al.* [10] designed an object-wise extension strategy based text detection framework, which only needed to predict shrink-masks by one segmentation header. Importantly, the object-wise strategy saved much computational cost compared with [9].

III. Methodology

In this section, we introduce the overall structure of the proposed ZTD firstly. Then, the details of Zoom Out Mod-

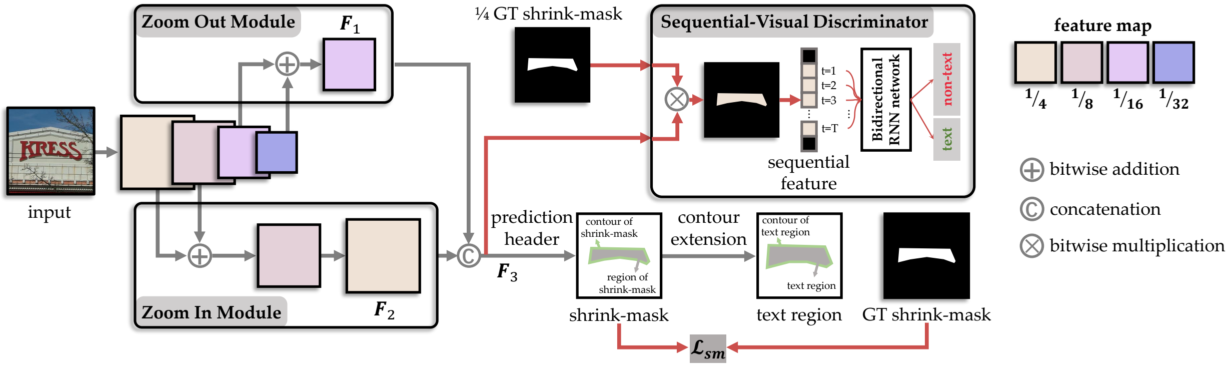


Fig. 2. Overall structure of the proposed Zoom Text Detector, which consists of Zoom Out Module, Zoom In Module, Sequential-Visual Discriminator, shrink-mask prediction header, and contour extension process. Red flows are training only operators.

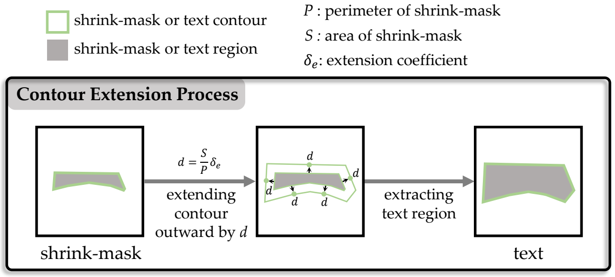

shrink-mask or text contour
shrink-mask or text region
P : perimeter of shrink-mask
S : area of shrink-mask
δ_e: extension coefficient
Contour Extension Process
$d = \frac{S}{P}\delta_e$
extending contour outward by d
d
extracting text region
shrink-mask
text

Fig. 3. Visualization of text contour extension process.

ule (ZOM), Zoom In Module (ZIM), and Sequential-Visual Discriminator (SVD) are described and shown through visualization. In the end, the optimization function is given.

A. Overall Pipeline

The architecture of ZTD is shown in Figure 2, which is composed of backbone, Zoom Out Module (ZOM), Zoom In Module (ZIM), Sequential-Visual Discriminator (SVD), prediction header, and contour extension process. For backbone, it is used for the generation of multi-level feature maps f_1, f_2, f_3, and f_4 corresponding to the image size of $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$, respectively. To avoid the phenomena of feature defocusing and detail loss that exist in current methods [9], [10], ZOM and ZIM are proposed. Specifically, as shown in Fig. 2, ZOM fuses f_3 and f_4 at first. Then, a coarse segmentation task is conducted on F_1 to help ZTD to extract semantic features, which promotes the discrimination of shrink-masks from the background. For ZIM, it utilizes fine features to force ZTD to recognize the margins, which facilitates our method to distinguish shrink-masks from the margins. Considering false-positive samples enjoy similar visual features (such as color, texture, and edge) with shrink-masks, SVD encourages ZTD to extract the sequential and visual features to distinguish them in the temporal and spatial domains. The header consists of two transposed convolution layers. It utilizes the hybrid features from ZOM, ZIM, and SVD to predict shrink-masks accurately. In the end, the texts can be rebuilt by the contour extension process (as shown in Fig. 3) directly, where the extension distance is computed by the formula in [42].

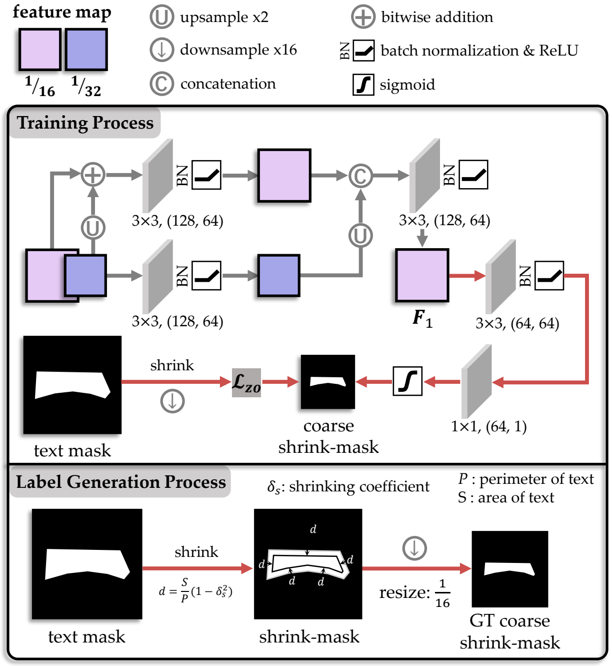


Fig. 4. Visualization of the structure and label generation process of Zoom Out Module. Red flows are training only operators

B. Zoom Out Module

According to [1], semantic information can strengthen discrimination of shrink-masks from the background. However, for recent lightweight methods [9], [10], the phenomenon of feature defocusing limits the extraction of semantic features, where coarse layers are optimized by the fine-grained supervision information. Considering the above problems, ZOM is proposed to provide coarse-grained optimization objectives for coarse layers to facilitate the extraction of semantic features.

The structure of ZOM is shown in Fig. 4. The module fuses two coarse features to strengthen the expression of f_4 firstly. To save computational cost, we use two 3 × 3 convolution layers to reduce the channels of feature maps in f_3 and f_4 branches, respectively. Considering the concatenation operator can provide a larger mapping space compared to point-wise addition and experimental results in [43] verify the effectiveness of the concatenation layer for segmentation tasks, f_3 and f_4 are concatenated in this structure. Following the design strategy of the lightweight network, we reduce

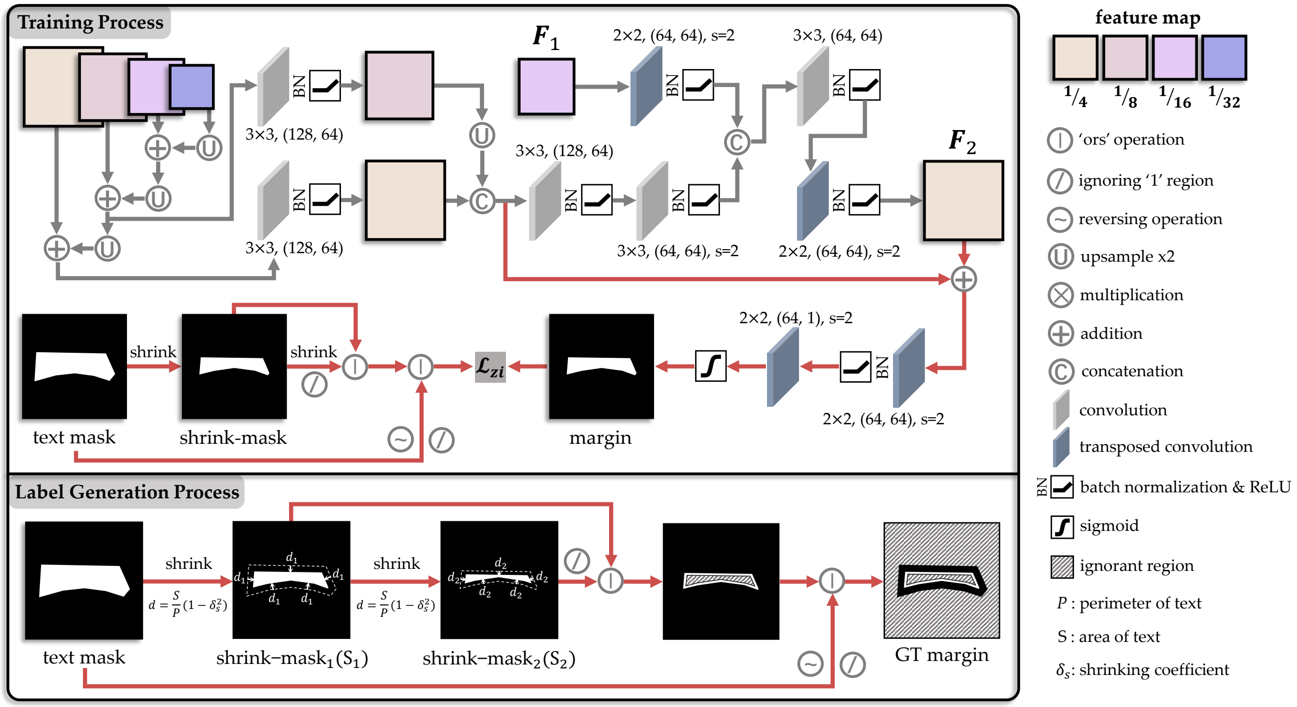


Fig. 5. Visualization of the structure and label generation process of Zoom In Module. Red flows are training only operators

the concatenated feature channels to half of the original to generate the feature map F_1.

In the training process, a 3×3 convolution is used for F_1 smoothing, and a 1×1 convolution is conducted on the smooth features to segment the coarse shrink-mask map. The shrink-mask can be obtained by performing a sigmoid function on the map. The loss function L_{zo} evaluates the error between the predicted coarse shrink-mask and the corresponding ground-truth. Furthermore, we visualize the label generation process of the coarse shrink-mask in Fig. 4. Specifically, text contour is shrunk inward by d that is computed by the formula in [42] and the region in shrunk contour is treated as shrink-mask. The ground-truth of coarse shrink-mask is obtained by resizing shrink-mask to $\frac{1}{16}$ of image size. Particularly, except the final 1×1 convolution, each convolutional layer is followed by a BatchNorm layer [44] and rectified linear unit (ReLU) [45].

C. Zoom In Module

The shrink-mask is generated by shrinking the text contour (as shown in Fig. 4). It means both shrink-masks and the corresponding margins belong to texts, which makes it hard to discriminate them. For current methods, they ignore the margins recognition (the phenomenon of detail loss), which leads to ambiguous shrink-mask edges and further results in text adhesion and miss detection. Considering fine-grained information is useful for discriminating shrink-masks and the margins, ZIM is designed to encourage ZTD to utilize the information to recognize the margins, which helps to predict shrink-mask edges precisely.

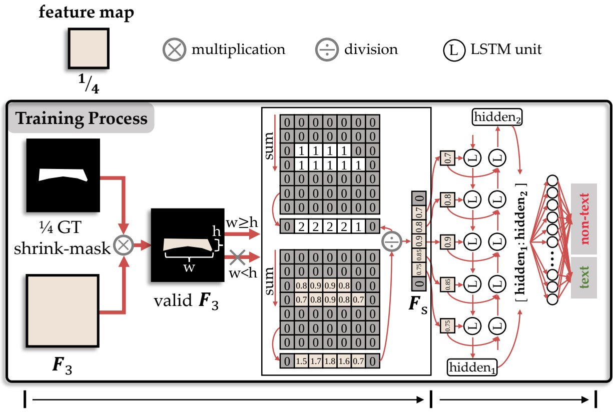
The architecture of ZIM is illustrated in Fig. 5. In the front part, it conducts the same operations on f_1 and f_2 like ZOM to f_3 and f_4. The concatenated feature is treated as an input for the following two branches. For the first one, the feature is downsampled by a 3×3 convolution with two strides after reducing channels. Then, it is concatenated with upsampled F_1 from ZOM, where F_1 is used to provide semantic information for enhancing fine-grained margin recognition. Next, we further upsample the feature by transposed convolution as $\frac{1}{4}$ size of image for pixel-wise segmentation. In the end, considering the image details are lost with the increase of network layers, the above concatenated feature is skip connected with the upsampled feature F_2.

In the training process, the combination of two transposed convolutions and one sigmoid function is used to predict the margins, which is helpful to strengthen the model's ability to recognize shrink-mask edges. The loss function L_{zi} evaluates the differences between the predicted margins and the corresponding ground-truth. As shown in Fig. 5, the label generation process includes five steps: 1) Shrinking text contour to obtain the shrink-mask S_1; 2) Shrinking the contour of S_1 to obtain a smaller shrink-mask S_2; 3) Ignoring the '1' region in S_2 and conducting 'ors' operation between S_1 and ignored S_2; 4) Reversing text mask and ignoring the '1' region; 5) Conducting 'ors' operation between the processed text mask and the result generated by step 3. The mentioned 'ors' operation is defined as:

$$\begin{gathered} p_{i,j}^a \text{ ors } p_{i,j}^b = ignorance, \\ \{p_{i,j}^a = ignorance \text{ or } p_{i,j}^b = ignorance\}, \end{gathered} \tag{1}$$

$$\begin{gathered} p_{i,j}^a \text{ ors } p_{i,j}^b = 1, \\ \{p_{i,j}^a = 1 \text{ or } p_{i,j}^b = 1\}, \\ \{p_{i,j}^a \neq ignorance \text{ and } p_{i,j}^b \neq ignorance\}, \end{gathered} \tag{2}$$

where $p_{i,j}^a$ and $p_{i,j}^b$ denote the pixel category of ith row and jth column on mask a and b, respectively.



(a) sequential feature pre-processing (b) discriminator

Fig. 6. Visualization of the structure of Sequential-Visual Discriminator. Red flows are training only operators.

D. Sequential-Visual Discriminator

False-positive samples enjoy highly similar visual features with shrink-masks (such as color, texture, and edge), which makes it difficult to discriminate them according to visual features only. Considering shrink-masks are equipped with sequential features, SVD is designed to encourage ZTD to utilize the combination of sequential and visual features to suppress false-positive samples.

Details of SVD are shown in Fig. 6. The module consists of sequential feature pre-processing and discriminator. The pre-processing takes $\frac{1}{4}$ GT shrink-mask and F_3 (shown in Fig. 2) as input. It first executes multiplication on them to generate valid F_3 and defines non-zero region as valid feature region. Then, the width w and height h of the region are computed, and the pixels of $\frac{1}{4}$ GT shrink-mask and valid F_3 are added along the column respectively to generate two vectors of them when $w \geq h$. In the end, the vector of valid F_3 is divided by the vector of $\frac{1}{4}$ GT shrink-mask to generate vector F_s. The discriminator is composed of an LSTM [46] based bidirectional RNN structure and a Fully Connect Network (FCN) based classifier. It abandons the zero regions of F_s and inputs the processed F_s into bidirectional RNN to extract two sequential features hidden_1 and hidden_2. The classifier treats the concatenation of hidden_1 and hidden_2 as input to estimate whether the region is shrink-mask, which encourages ZTD to extract the sequential feature and to combine it with the visual feature to suppress false-positive samples. Particularly, SVD is a training-only module, which brings no extra computational cost to the inference process and can be integrated into other detectors seamlessly.

E. Loss Function

As we can see from Fig. 2, the proposed ZTD is composed of shrink-mask prediction header, ZIM, ZOM, and SVD. Therefore, the overall objective function consists of $\mathcal{L}_{sm}$, $\mathcal{L}_{zi}$, $\mathcal{L}_{zo}$, and $\mathcal{L}_{svd}$, which can be formulated as:

$$\mathcal{L} = \alpha\mathcal{L}_{sm} + \beta\mathcal{L}_{zi} + \gamma\mathcal{L}_{zo} + \eta\mathcal{L}_{svd}, \tag{3}$$

where the parameters α, β, γ, and η balance the importance of different loss functions. They are set to 1, 0.25, 0.25, and 0.25 in the following experiments, respectively.

Optimization of shrink-mask prediction header. Dice loss [47] is proposed to evaluate the similarity of different binary masks. Particularly, it performs better than other loss functions when positive and negative samples are imbalanced, which is suitable for the shrink-mask prediction task. Therefore, we adopt the dice loss to evaluate the loss $\mathcal{L}_{sm}$ of this header, which is defined as:

$$\mathcal{L}_{sm} = 1 - \frac{2 \times |\mathrm{SM_p} \cap \mathrm{SM_g}| + \varepsilon}{|\mathrm{SM_p}| + |\mathrm{SM_g}| + \varepsilon}, \tag{4}$$

where $\mathrm{SM_p}$ and $\mathrm{SM_g}$ indicate the predicted shrink-mask and the corresponding ground-truth. Considering that there may be no positive samples in ground-truth, we set ε as 1 to avoid the denominator equal to 0.

Optimization of Zoom In Module. As we mentioned before, the shrink-mask is generated by shrinking the text contour inward by a specific distance, which means both the margins and shrink-masks are parts of texts. It makes existing methods hard to discriminate them, which may lead to ambiguous shrink-mask edges and further influence model performance. To recognize the edges accurately, we focus on the margins through ZIM. The loss function $\mathcal{L}_{zi}$ of ZIM can be expressed as:

$$\mathcal{L}_{zi} = 1 - \frac{2 \times |(\mathrm{ZI_p}) \cap \mathrm{ZI_g}| + \varepsilon}{|(\mathrm{ZI_p})| + |\mathrm{ZI_g}| + \varepsilon}, \tag{5}$$

where $\mathrm{ZI_p}$ denotes the predicted binary mask of the margin and $\mathrm{ZI_g}$ is the corresponding label.

Optimization of Zoom Out Module. The label of this module is shrink-mask of $\frac{1}{16}$ stride, which can be generated by the Label Generation Process in Fig. 4. The same as shrink-mask prediction header, dice loss is used for the evaluation of the loss $\mathcal{L}_{zo}$ between the predicted binary mask and ground-truth:

$$\mathcal{L}_{zo} = 1 - \frac{2 \times |\mathrm{ZO_p} \cap \mathrm{ZO_g}| + \varepsilon}{|\mathrm{ZO_p}| + |\mathrm{ZO_g}| + \varepsilon}, \tag{6}$$

where $\mathrm{ZO_p}$ and $\mathrm{ZO_g}$ are the predicted coarse shrink-mask and the corresponding ground-truth.

Optimization of Sequential-Visual Discriminator. Considering false-positive samples enjoy similar visual features with texts, SVD is presented to encourage our model to suppress them by the combination of sequential and visual features. For this classification task, we adopt BCE loss to measure the loss $\mathcal{L}_{svd}$ of this module:

$$\mathcal{L}_{svd} = -(\mathrm{S_p} \times \log(\mathrm{S_g}) + (1 - \mathrm{S_p}) \times \log(1 - \mathrm{S_g}), \tag{7}$$

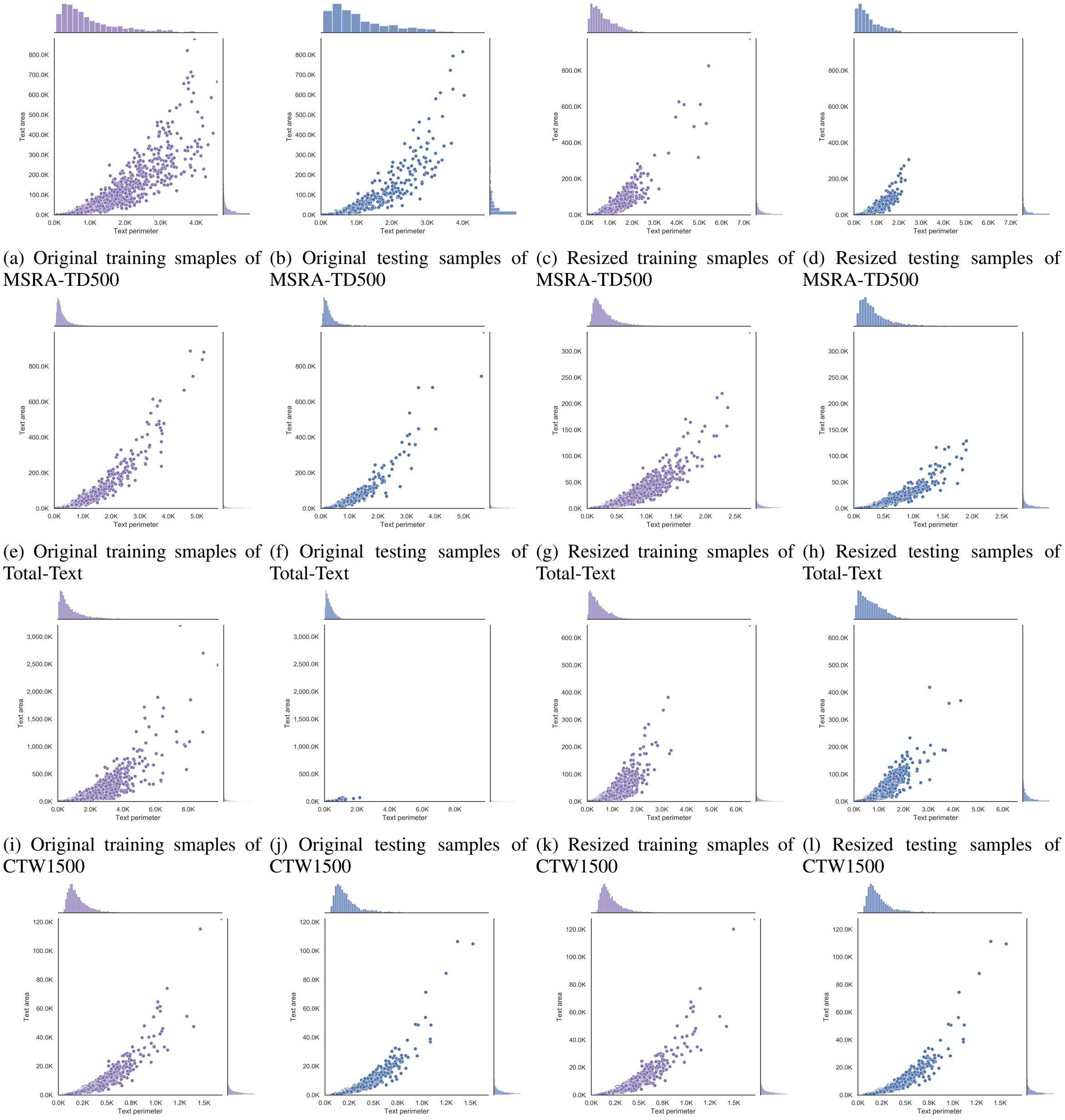
where $\mathrm{S_p}$ is the probability whether the region is shrink-mask and $\mathrm{S_g}$ denotes the ground-truth.

IV. Experiments

A. Datasets

To verify the effectiveness and robustness of our method to the texts with different shapes, scales, and aspect ratios, we evaluate ZTD on the four representative public datasets:

MSRA-TD500 [48] is a dataset consisting of line-level Chinese and English text instances. There are 300 training



(a) Original training smaples of MSRA-TD500 (b) Original testing samples of MSRA-TD500 (c) Resized training smaples of MSRA-TD500 (d) Resized testing samples of MSRA-TD500

(e) Original training smaples of Total-Text (f) Original testing samples of Total-Text (g) Resized training smaples of Total-Text (h) Resized testing samples of Total-Text

(i) Original training smaples of CTW1500 (j) Original testing samples of CTW1500 (k) Resized training smaples of CTW1500 (l) Resized testing samples of CTW1500

(m) Original training smaples of ICDAR2015 (n) Original testing samples of ICDAR2015 (o) Resized training smaples of ICDAR2015 (p) Resized testing samples of ICDAR2015

Fig. 7. Visualization of geometry characteristics of text instances in MSRA-TD500 (the first row), Total-Text (the second row), CTW1500 (the third row), and ICDAR2015 (the fourth row) datasets.

images and 200 testing images, respectively. The same as previous works, we introduce 400 extra images from HUST-TR400 [49] as training data.

Total-Text [50] contains horizontal, multi-oriented, curved and other irregular-shaped texts. Except for English texts, there are still some Chinese and Japanese samples, which brings difficulty for detection. This dataset contains 1255 training images and 300 testing images, respectively.

CTW1500 [51] has 1000 training images and 500 testing images. Different from Total-Text, this dataset mainly consists of line-level arbitrary-shaped text instances.

ICDAR2015 [52] is proposed in ICDAR 2015 Robust Reading Competition. Compared with the above three public benchmarks, ICDAR2015 has a more complicated background, which makes it hard to distinguish text and interference region. The same as CTW1500, ICDAR2015 utilizes 1000 images to train the model and 500 images to evaluate the detection performance.

The geometry characteristics of text instances of different datasets are shown in Fig. 7. For the original datasets (as illustrated in the first and second columns in Fig. 7), text scales of CTW1500 are almost 25 times bigger than the text of ICDAR2015. Moreover, there are huge characteristic differences between the training and testing text instances of CTW1500, which brings difficulty for text detection. For the text instances of MSRA-TD500 and Total-Text, they enjoy

TABLE I
DETECTION RESULTS OF ZTD WITH DIFFERENT SETTINGS ON MSRA-TD500. "S: 736" MEANS THAT THE SHORT SIDE OF EACH TESTING IMAGE IS RESIZED TO BE 736 PIXELS. "BASELINE" MEANS THE FRAMEWORK EQUIPPED WITH SHRINK-MASK PREDICTION HEADER ONLY. "EXT." INDICATES THAT ZTDL IS PRE-TRAINED ON SYNTHTEXT [53].

Image scale for testing (S : 736)									
#	Methods	ZIM	ZOM	SVD	Ext.	P	R	F	FPS
1	baseline					86.4	79.9	83.0	64.1
2	baseline+	✓				87.5	80.5	83.9	64.1
3	baseline+	✓	✓			90.7	80.3	85.2	59.2
4	baseline+	✓	✓	✓		92.2	80.9	86.2	59.2
5	baseline+	✓	✓	✓	✓	91.6	82.4	86.8	59.2

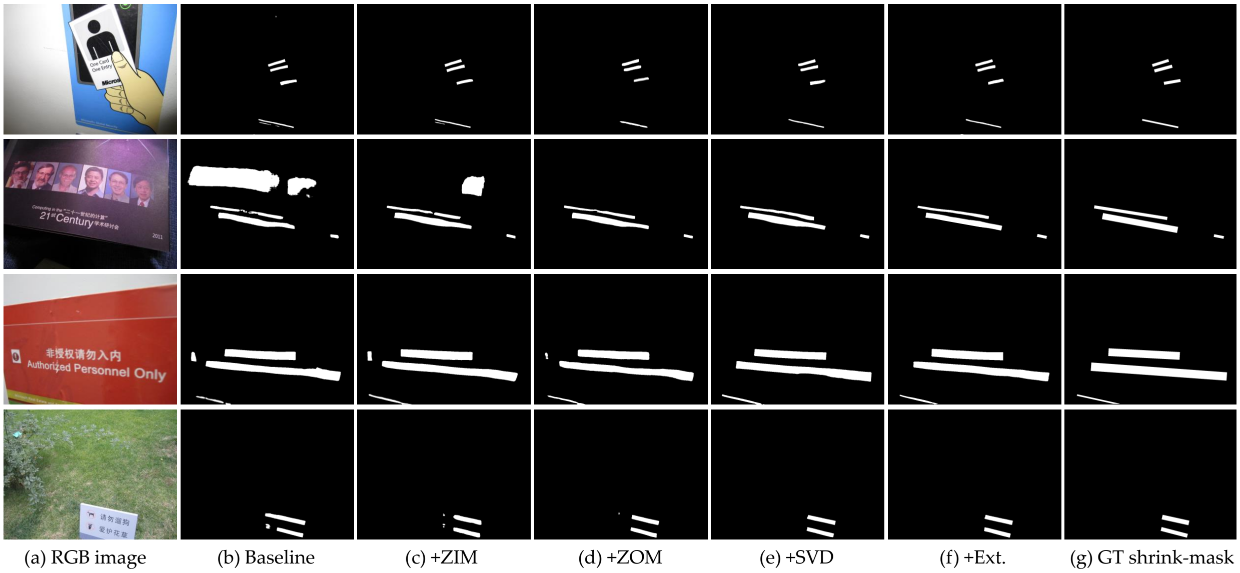

(a) RGB image (b) Baseline (c) +ZIM (d) +ZOM (e) +SVD (f) +Ext. (g) GT shrink-mask

Fig. 8. Visualization of the predicted shrink-masks of ZTD with different settings.

similar geometry characteristics. To ensure a fair comparison environment, we resize the short sides of original images to specific sizes to generate resized text instances. It can be found from the third and fourth columns in Fig. 7 that the resized training texts and testing texts enjoy similar geometry characteristic distributions.

B. Implementation Details

The overall architecture of our method is shown in Fig. 2, where the feature maps (f_1, f_2, f_3, and f_4) behind input are generated by different stages (stage1, stage2, stage3, and stage4 respectively) of ResNet-18 [54].

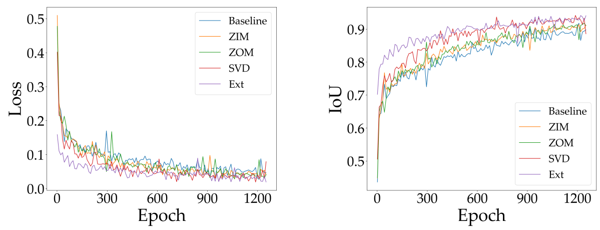
In the data pre-processing stage, the training samples are increased by the following augmentation strategies: (1) random scaling (including image size and aspect); (2) random horizontal flipping; (3) random rotating in the range of (-10, 10); (4) random cropping and padding.

In the initializing stage, the backbone of ZTD is pre-trained on ImageNet [55] and the rest of the layers are initialized by the strategy proposed in [56]. In the training process, the Adam [57] is deployed to optimize the model. For learning rate, it is initialized as 0.001 and adjusted through 'polylr' strategy. In the following experiments, our model is pre-trained on the SynthText dataset for 1 epoch and finetuned on the corresponding real-world datasets for 1200 epochs. The training batch size is set to 16. Moreover, the text instances labeled as DO NOT CARE are ignored during both training and testing stages. In the inference process, the red flows in Fig. 2, Fig. 4, Fig. 5, and Fig. 6 are abandoned, which is helpful to facilitate detection speed. All the experiments in this paper are performed on a workstation with 1080Ti GPU.

C. Ablation Study

To verify the effectiveness of the proposed ZIM, ZOM, and SVD, we conduct an ablation study in this section. Furthermore, we explore the impacts of each sub-loss of $\mathcal{L}$ and the importances of different RNN units of SVD, respectively. The details of experimental results are described in the following paragraphs.

Effectiveness of Zoom In Module. As described in Section III-A, text contour is generated through extending shrink-mask contour outward by a specific distance, which means the accuracy of shrink-mask edge influences model performance directly. ZIM is proposed to force our method to focus on the margins, which helps ZTD to recognize shrink-mask edges precisely. Compared with baseline (Table I #1), ZIM brings 0.9% improvements in F-measure. Particularly, it brings no extra computational cost to the inference process, which benefits from the sharing structure between baseline and ZIM. Moreover, as we can see from the second and third columns

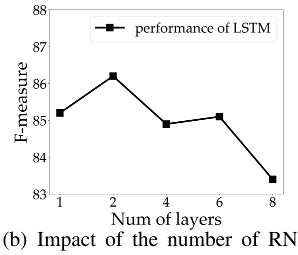


(a) Details of training loss (b) Details of training IoU

Fig. 9. Convergence analysis of ZTD with different settings in Table I on MSRA dataset. 'IoU' means the Intersection of Union between predicted shrink-mask and the corresponding ground-truth.

Unit	P	R	F
RNN	89.7	81.4	85.3
GRU	90.5	81.5	85.8
LSTM	92.2	80.9	86.2

(a) Impact of different RNN units for detection accuracy

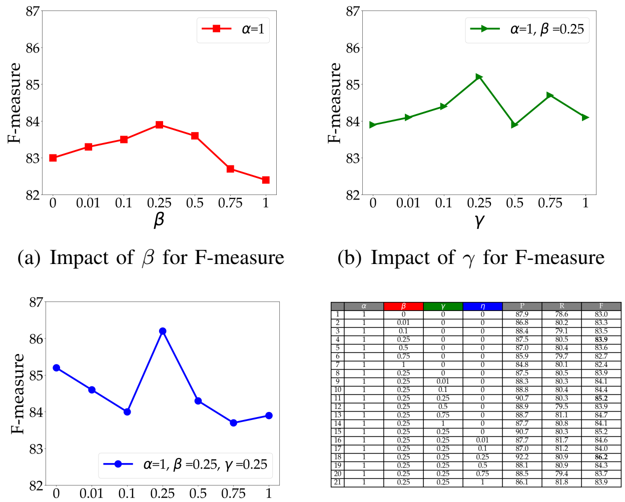


(b) Impact of the number of RNN layers for detection accuracy

Fig. 10. Detection results of SVD with different settings on MSRA-TD500. 'Unit' indicates the unit of RNN structure in SVD.

in Fig. 8, ZIM encourages the baseline to perform better for the recognition of shrink-mask edges. The loss curve of the training process of baseline+ZIM is shown in Fig. 9, which enjoys a faster convergence speed compared to baseline.

Effectiveness of Zoom Out Module. As we mentioned before, ZOM is presented to avoid the phenomenon of feature defocusing to enhance the discrimination of shrink-masks from the background. It is found in Table I, the #3 model outperforms baseline 2.2% and #2 1.3% in F-measure, respectively, which verifies the effectiveness of ZOM. Meanwhile, we compare the detection results of the #3 model with baseline and baseline+ZIM in Fig. 8. It can be seen that ZOM helps our method to discriminate shrink-masks from some interference regions of the background effectively.



(a) Impact of β for F-measure (b) Impact of γ for F-measure

(c) Impact of η for F-measure

	α	β	γ	η	P	R	F
1	1	0	0	0	87.9	78.6	83.0
2	1	0.01	0	0	86.8	80.2	83.3
3	1	0.1	0	0	88.4	79.1	83.5
4	1	0.25	0	0	87.5	80.5	**83.9**
5	1	0.5	0	0	87.0	80.4	83.6
6	1	0.75	0	0	85.9	79.7	82.7
7	1	1	0	0	84.8	80.1	82.4
8	1	0.25	0	0	87.5	80.5	83.9
9	1	0.25	0.01	0	88.3	80.3	84.1
10	1	0.25	0.1	0	88.8	80.4	84.4
11	1	0.25	0.25	0	90.7	80.3	**85.2**
12	1	0.25	0.5	0	88.9	79.5	83.9
13	1	0.25	0.75	0	88.7	81.1	84.7
14	1	0.25	1	0	87.7	80.8	84.1
15	1	0.25	0.25	0	90.7	80.3	85.2
16	1	0.25	0.25	0.01	87.7	81.7	84.6
17	1	0.25	0.25	0.1	87.0	81.2	84.0
18	1	0.25	0.25	0.25	92.2	80.9	**86.2**
19	1	0.25	0.25	0.5	88.1	80.9	84.3
20	1	0.25	0.25	0.75	88.5	79.4	83.7
21	1	0.25	0.25	1	86.1	81.8	83.9

(d) Impact details of β, γ, and η

Fig. 11. Ablation study for the impact of β, γ, and η on performance.

Effectiveness of Sequential-Visual Discriminator. Considering false-positive samples enjoy highly similar visual features with shrink-masks and are hard to recognize according to visual features only, SVD is designed to encourage ZTD to suppress them by the combination of sequential and visual features. As shown in the fourth and fifth columns in Fig. 8, we can find that SVD helps our method to suppress false-positive samples effectively. The experimental results in Table I #4 also verify the effectiveness of SVD. Moreover, we pre-train our model on SynthText in this paper to keep a fair comparison environment with existing methods. As shown in Fig. 8 (f), pre-training our model on SythText improve the accuracy of predicted shrink-masks and brings 0.6% F-measure (Table I #5). Furthermore, we can see from Fig. 9 (a), the pre-trained model converges more faster than others.

Impacts of Different Settings for SVD. SVD extracts sequential features by RNN and inputs the features into FCN-based classifier to help ZTD to discriminate shrink-masks from false-positive samples (as shown in Fig. 6). In this section, we explore the influences of different units and the number of RNN layers for the extraction of sequential features. As shown in Fig. 10 (a), LSTM-based SVD brings 0.9% and 0.4% improvements in F-measure compared to normal RNN unit and GRU, respectively. Moreover, we can see from Fig. 10 (b), ZTD achieves the optimal performance when the number of RNN layers is set to 2. The above experimental results not only verify the positive effect of sequential features for the discrimination of shrink-masks but also demonstrate the prominent performance of LSTM to extract the sequential features of very long shrink-masks.

Importances of Different Sub-losses. As described in Section III-E, the optimization function $\mathcal{L}$ is composed of $\mathcal{L}_{sm}$, $\mathcal{L}_{zi}$, $\mathcal{L}_{zo}$, and $\mathcal{L}_{svd}$. α, β, γ, and η are the corresponding weights. In this section, we tune the value of a single weight and keep others fixed to evaluate the importance of each sub-loss. All experimental results are shown in Fig. 11. α is the weight of shrink-mask prediction header, it is set to 1 empirically. For β, γ, and η, we first analyze the importance of β. As shown in Fig. 11 (a), the proposed ZTD achieves the optimal performance when β is equal to 0.25, which indicates ZIM has a certain positive effect for the prediction of the shrink-mask. Furthermore, we perform the same analysis for γ. As demonstrated in Fig. 11 (b), the model performance is always better than baseline+ZIM when tuning γ in the range of 0–1, which demonstrates the effectiveness of ZOM for the distinguishment between shrink-masks and the background. Moreover, we test η by the same experiment. As shown in Fig. 11 (c), ZTD achieves the best performance when η is set to 0.25 and the performance fluctuates when η is close to 0.1 and 0.75. In Fig. 11 (d), the impact details of β, γ, and η on model performance are described, which helps to understand the impartances of different sub-losses intuitively.

TABLE II
PERFORMANCE COMPARISON ON MSRA-TD500 DATASET.

Methods	P	R	F	FPS
Accuracy Prior				
PixelLink [30] (AAAI 2018)	83.0	73.2	77.8	-
RRD [25] (CVPR 2018)	87.0	73.0	79.0	10
CRAFT [33] (CVPR 2018)	88.2	78.2	82.9	8.6
SAE [38] (CVPR 2019)	84.2	81.7	82.9	-
TexrField [39] (TIP 2019)	87.4	75.9	81.3	-
OPMP [7] (TMM 2020)	86.0	83.4	84.7	1.6
SAVTD [5] (CVPR 2021)	89.2	81.5	85.2	-
GV [26] (TPAMI 2021)	88.8	84.3	86.5	15.0
Comprehensive Performance Prior				
DB [10] (AAAI 2020)	90.4	76.3	82.8	62.0
PAN [9] (ICCV 2019)	84.4	83.8	84.1	30.2
PAN++ [40] (TPAMI 2021)	85.3	84.0	84.7	32.5
ZTD-512 (Ours)	90.5	82.1	86.1	97.4
ZTD-640 (Ours)	91.5	81.6	86.3	72.7
ZTD-736 (Ours)	91.6	82.4	86.8	59.2

D. Comparison with State-of-the-Art Methods

To verify the superior performance of ZTD, we compare it with the existing competitors on multiple representative public benchmarks (such as MSRA-TD500, Total-Text, CTW1500, and ICDAR2015) in this section. Considering existing text detection methods can be categorized into accuracy prior and comprehensive performance prior methods roughly (as mentioned in Section II), we analyze the advantages of ZTD over them respectively in the following experiments.

Evaluation on MSRA-TD500. We evaluate the performance of ZTD for detecting multi-language long straight text instances on MSRA-TD500 dataset. The experimental results are shown in Table II. It is found that our method outperforms existing state-of-the-art (SOTA) approaches in both detection accuracy and speed. Specifically, for GV [26], the best accuracy prior method, ZTD-736 outperforms it by 0.3% in F-measure. It is because the proposed ZIM, ZOM, and SVD enhance the model's ability to recognize shrink-masks. Meanwhile, benefiting from the lightweight CNN model and simple post-processing, our method runs 4 times faster than it. Furthermore, the comprehensive performance of ZTD-736 outperforms PAN [9], PAN++ [30] a lot. Though DB [10] achieves 62.0 FPS in detection speed, ZTD-512 is 35.4 FPS faster than it. Some qualitative results are shown in Fig. 12 (a). The experiments on MSTA-TD500 demonstrate the effectiveness of ZTD for detecting long text instances, even they are multilingual.

Evaluation on Total-Text. To verify the robustness of ZTD to detect word-level irregular-shaped texts, we evaluate it on Total-Text benchmark. The same as the experimental conclusion on MSRA-TD500, our method is superior to others in both detection accuracy and speed. As shown in Table III, for accuracy prior methods, MaskTextSpotter [13], ContourNet [35], and DRRG [31] achieve 85.2%, 85.4%, and 85.7% in F-measure, respectively. For comprehensive performance prior approaches, PAN [9] and PAN++ [30] enjoy comparable detection accuracy with accuracy prior methods. DB [10] performs better in detection speed. Compared with the above methods, ZTD can achieve 86.0% in F-measure and 75.2 FPS. Since many texts are close to each other in Total-Text, existing methods are hard to separate them efficiently. Benefiting from ZIM, our method enjoys a strong ability to recognize shrink-mask edges, which helps ZTD to avoid text adhesion problem. We further display some detection results in Fig. 12 (b). It is found that adhesive texts are separated successfully.

TABLE III
PERFORMANCE COMPARISON ON TOTAL-TEXT DATASET.

Methods	P	R	F	FPS
Accuracy Prior				
TextSnake [32] (ECCV 2018)	82.7	74.5	78.4	-
TextDragon [58] (ICCV 2019)	85.6	75.7	80.3	-
TextField [39] (TIP 2019)	81.2	79.9	80.6	-
Boundary [36] (AAAI 2020)	85.2	83.5	84.3	-
ContourNet [35] (CVPR 2020)	86.9	83.9	85.4	3.8
DRRG [31] (CVPR 2020)	86.5	84.9	85.7	-
FCENet [11] (CVPR 2021)	87.4	79.8	83.4	-
ReLaText [12] (PR 2021)	84.8	83.1	84.0	-
MaskTextSpotter [13] (TPAMI 2021)	88.3	82.4	85.2	-
Comprehensive Performance Prior				
DB [10] (AAAI 2020)	88.3	77.9	82.8	50.0
PAN [9] (ICCV 2019)	89.3	81.0	85.0	39.6
PAN++ [40] (TPAMI 2021)	89.9	81.0	85.3	38.3
KPN [4] (TNNLS 2022)	88.0	82.3	85.1	22.7
ZTD-512 (Ours)	90.5	80.6	85.3	93.2
ZTD-640 (Ours)	90.1	82.3	86.0	75.2

TABLE IV
PERFORMANCE COMPARISON ON CTW1500 DATASET.

Methods	P	R	F	FPS
Accuracy Prior				
CRAFT [33] (CVPR 2018)	86.0	81.1	83.5	-
LOMO [17] (CVPR 2019)	85.7	76.5	80.8	-
TexrField [39] (TIP 2019)	83.0	79.8	81.4	-
OPMP [7] (TMM 2020)	85.1	80.8	82.9	1.4
TextRay [27] (ACMMM 2020)	82.8	80.4	81.6	-
DRRG [31] (CVPR 2020)	85.9	83.0	84.5	-
FCENet [11] (CVPR 2021)	85.7	80.7	83.1	-
ReLaText [12] (PR 2021)	86.0	83.3	84.8	10.6
Comprehensive Performance Prior				
DB [10] (AAAI 2020)	84.8	77.5	81.0	55.0
PAN [9] (ICCV 2019)	86.4	81.2	83.7	39.8
PAN++ [40] (TPAMI 2021)	87.1	81.1	84.0	36.0
KPN [4] (TNNLS 2022)	84.0	82.9	83.4	24.3
ZTD-640 (Ours)	88.4	80.2	84.1	76.9

Evaluation on CTW1500. Experiments on CTW1500 show the effectiveness of ZTD for detecting line-level arbitrary-shaped texts. All experimental results are shown in Table IV. Our method runs 76.9 FPS and is faster than other methods at least by 21.9 FPS. The outstanding detection speed benefits from the efficient text representation method and lightweight CNN model. Specifically, PAN [9] and PAN++ [30] reconstruct text contours through pixel-wise extension strategy. Unlike these comprehensive performance prior methods, ZT-D adopts an object-wise extension strategy. Moreover, our method optimizes the CNN model to design a lightweight and efficient network, which significantly gains our approach in detection speed. For accuracy prior methods, though the

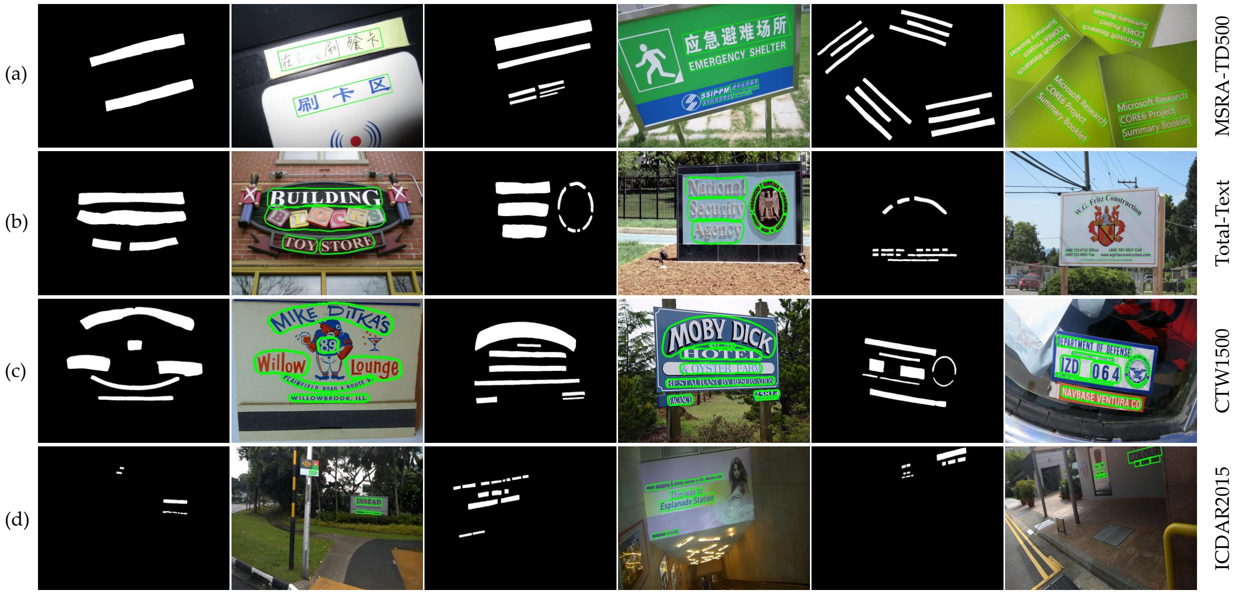


Fig. 12. Visualization of some qualitative detection results of ZTD on MSRA-TD500, Total-Text, CTW1500, and ICDAR2015 datasets. Binary masks are the predicted shrink-masks and RGB images show the rebuilt text contours based on the predicted shrink-masks.

TABLE V
PERFORMANCE COMPARISON ON ICDAR2015 DATASET.

Methods	P	R	F	FPS
Accuracy Prior				
TextSnake [32] (ECCV 2018)	84.9	80.4	82.6	-
CornerNet [34] (ECCV 2018)	89.5	79.7	84.3	1.0
PSE [41] (CVPR 2019)	86.9	84.5	85.7	1.6
SAE [38] (CVPR 2019)	85.1	84.5	84.8	-
Boundary [36] (AAAI 2020)	88.1	82.2	85.0	-
FCENet [11] (CVPR 2021)	85.1	84.2	84.6	-
TEETS [24] (TPAMI 2021)	-	-	85.0	-
MaskTextSpotter [13] (TPAMI 2021)	85.8	81.2	83.4	4.8
PolarMask++ [37] (TPAMI 2021)	87.3	83.5	85.4	10.0
Comprehensive Performance Prior				
DB [10] (AAAI 2020)	86.8	78.4	82.3	48.0
PAN [9] (ICCV 2019)	84.0	81.9	82.9	26.1
PAN++ [40] (TPAMI 2021)	85.9	80.4	83.1	28.2
ZTD-736 (Ours)	87.5	79.0	83.0	48.3

TABLE VI
CROSS-DATASET EVALUATIONS ON WORD-LEVEL AND LINE-LEVEL DATASETS.

Type	Traning dataset	Test dataset	P	R	F
word-level	ICDAR2015	Total-Text	78.5	64.1	70.6
	Total-Text	ICDAR2015	79.8	69.3	74.2
Line-level	MSRA-TD500	CTW1500	84.1	73.4	78.4
	CTW1500	MSRA-TD500	86.8	77.9	82.1

detection accuracy of ZTD is not as well as some methods (such as DRRG [31] and ReLaText [12]), the proposed approach has at least 7 times faster speed than them. As shown in Fig. 12 (c), the above experiments and the visualization of detection results demonstrate the effectiveness of ZTD to recognize long irregular-shaped text instances.

Evaluation on ICDAR2015. To verify the robustness of ZTD to detect multi-oriented text instances from the complicated background, we compare ZTD with existing text detection methods on ICDAR 2015 benchmark. As exhibited in Table V, our method achieves 83.0% F-measure with 48.3 FPS, which outperforms DB [10] and PAN [9] in both detection accuracy and speed. Moreover, ZTD can run 2 times faster than PAN++ [30] and achieves comparable detection accuracy to it. Compared with the accuracy prior methods, the proposed detector keeps considerable superiority in detection speed and accomplishes the best balance between detection accuracy and speed. The superior comprehensive performance brings great potential for a wide range of applications. The results in Table V and Fig. 12 (d) demonstrate our method can recognize the texts with various scales and multi-orientations from the complex background effectively.

E. Cross Dataset Text Detection

We conduct multiple comparison experiments in Section IV-D and show the superior performance in both detection accuracy and speed of our method. To further verify the generalization performance of ZTD, we further evaluate it through a series of cross-train-test experiments. Specifically, considering ICDAR2015 and Total-Text are word-level datasets, MSRA-TD500 and CTW1500 are line-level benchmarks, we design two sets of experiments on word-level and line-level datasets, respectively. At first, our method is trained on the training images of ICDAR2015 and Total-Text. Then, we evaluate ZTD on the testing images of Total-Text and ICDAR2015. As we can see from Table VI, ZTD achieves 70.6% and 74.2% in F-measure, respectively. For line-level datasets, the same cross-train-test experiments are conducted. Particularly, our method achieves 82.1% in F-measure when it is trained on CTW1500 and tested on MSRA-TD500, which surpasses many methods (e.g., PixelLink [30], RRD [25], and TextField [39]) that is trained on MSRA-TD500, which shows the effectiveness of

TABLE VII
TIME CONSUMPTION OF ZTD ON FOUR PUBLIC BENCHMARKS. DIFFERENT STAGES OF ZTD INCLUDE BACKBONE, ZOOM (ZIM AND ZOM), HEAD (SHRINK-MASK PREDICTION HEADER), AND POST (CONTOUR EXTENSION PROCESS). 'IMAGE SIZE' DENOTES THE SIZE OF THE IMAGE SHORT SIDE.

Datasets	Image size	Time consumption (ms)				FPS	F
		Backbone	Zoom	Head	Post		
MSRA-TD500	736	8.0	4.2	3.3	1.4	59.2	86.8
Total-Text	640	6.2	3.1	2.6	1.4	75.2	86.0
CTW1500	640	6.0	3.1	2.5	1.4	76.9	84.1
ICDAR2015	736	9.9	5.1	4.1	1.6	48.3	83.0

our method for long text detection and the generalization performance in different scenes.

F. Speed Analysis

The above experiments demonstrate the outstanding comprehensive performance of our method. Especially in terms of detection speed, the proposed ZTD enjoys an obvious advantage compared to previous algorithms. To verify the high efficiency of the designed framework, we analyze the time consumption details of ZTD's different stages in this section. The experimental details as described in Table VII. To keep a fair comparison environment, we resize the short sides of images as 736, 640, 640, and 736 for MSRA-TD500, Total-Text, CTW1500, and ICDAR2015, respectively. It is found that 'Backbone' takes about half of the total time. It is mainly because 'Backbone' is composed of plenty of convolution layers. Unlike 'Backbone', 'Zoom' and 'Head' are composed of fewer convolution layers. However, as a decoder structure, 'Head' needs to upsample feature maps to image size, which increases the time consumption. Therefore, though the layers of 'Zoom' and 'Head' are less than 'Backbone', they almost consume the same computational cost as 'Backbone'. The 'Post' denotes the contour extension process (shown in Fig. 3). Since it is an object-wise operation, the time consumption is much less than the above stages and does not influenced by the image size (as the comparison between different datasets in Table VII). The lightweight CNN model and object-wise contour extension process bring significant improvements for our method in detection speed, and the experimental results demonstrate this conclusion.

G. Failuer Cases

We have verified the superiority of the proposed ZTD in both detection accuracy and speed on multiple public benchmarks before. To further analyze the limitation of the proposed detector, we show some incorrect detection results. As demonstrated in Fig. 13, there are three challenging samples from ICDAR2015, CTW1500, and Total-Text datasets, respectively. For the sample from ICDAR2015, two text instances are missed detection, where blurred, and low color contrast are the failure reasons. For line-level (CTW1500) and word-level (Total-Text) datasets, half detection and overdetection are the current main problems, respectively. The above issues make there is still much room to improve the proposed method.

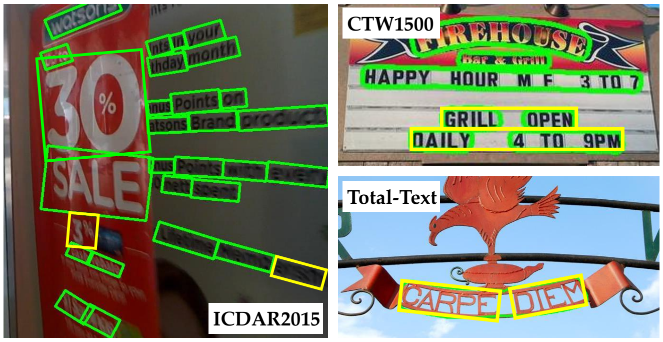


Fig. 13. Challenging samples from ICDAR2015, CTW1500, and Total-Text datasets. The green bounding boxes are the detection results from our method. The yellow ones are labels.

V. CONCLUSION

In this paper, we propose an efficient text detector inspired by the zoom process of the camera, termed as Zoom Text Detector (ZTD). By simulating the zooming out process of the camera, the detector can extract strong expressive semantic features from coarse layers, which enhances ZTD's ability to discriminate shrink-masks from the background significantly. Moreover, simulating the zooming in process of the camera encourages our method to focus on the margins, which helps to recognize shrink-mask edges accurately and avoid many problems (e.g., text adhesion and missed detection). Additionally, sequential features are extracted and combined with visual features to facilitate the presented approach to suppress false-positive samples effectively, which further improves the reliability of predicted shrink-masks. Extensive experiments show the effectiveness of ZOM, ZIM, and SVD. Comparisons on the multiple benchmarks demonstrate the superior comprehensive performance in both detection accuracy and speed of ZTD, which shows the great potential for a wide range of applications.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3431–3440.

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

[3] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.

[4] S.-X. Zhang, X. Zhu, J.-B. Hou, C. Yang, and X.-C. Yin, "Kernel proposal network for arbitrary shape text detection," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[5] W. Feng, F. Yin, X. Zhang, and C. Liu, "Semantic-aware video text detection," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 1695–1705.

[6] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, and L. Wang, "Abcnet: Real-time scene text spotting with adaptive bezier-curve network," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 9809–9818.

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

[8] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "East: an efficient and accurate scene text detector," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 5551–5560.

[9] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu, G. Yu, and C. Shen, "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 8440–8449.

[10] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time scene text detection with differentiable binarization." in *Proc. AAAI*, 2020, pp. 11 474–11 481.

[11] Y. Zhu, J. Chen, L. Liang, Z. Kuang, L. Jin, and W. Zhang, "Fourier contour embedding for arbitrary-shaped text detection," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 3123–3131.

[12] C. Ma, L. Sun, Z. Zhong, and Q. Huo, "Relatext: exploiting visual relationships for arbitrary-shaped scene text detection with graph convolutional networks," *Pattern Recognition*, vol. 111, p. 107684, 2021.

[13] M. Liao, P. Lyu, M. He, C. Yao, W. Wu, and X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 532–548, 2021.

[14] 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 Inform. Process. Syst.*, 2015, pp. 91–99.

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

[16] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, 2018.

[17] C. Zhang, B. Liang, Z. Huang, M. En, J. Han, E. Ding, and X. Ding, "Look more than once: An accurate detector for text of arbitrary shapes," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 10 552–10 561.

[18] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2550–2558.

[19] S. Tian, S. Lu, and C. Li, "Wetext: Scene text detection under weak supervision," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 1492–1500.

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg, "Ssd: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[21] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," *arXiv preprint arXiv:1611.06779*, 2016.

[22] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 3047–3055.

[23] 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, 2018.

[24] P. Wang, H. Li, and C. Shen, "Towards end-to-end text spotting in natural scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

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

[26] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G.-S. Xia, and X. Bai, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, 2020.

[27] F. Wang, Y. Chen, F. Wu, and X. Li, "Textray: Contour-based geometric modeling for arbitrary-shaped scene text detection," in *ACM. Multimedia*, 2020, pp. 111–119.

[28] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "Densebox: Unifying landmark localization with end to end object detection," *arXiv preprint arXiv:1509.04874*, 2015.

[29] X. Li, M. Chen, F. Nie, and Q. Wang, "A multiview-based parameter free framework for group detection," in *Proc. AAAI*, 2017, pp. 4147–4153.

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

[31] S. Zhang, X. Zhu, J. Hou, C. Liu, C. Yang, H. Wang, and X. Yin, "Deep relational reasoning graph network for arbitrary shape text detection," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 9696–9705.

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

[33] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 9365–9374.

[34] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 734–750.

[35] Y. Wang, H. Xie, Z. Zha, M. Xing, Z. Fu, and Y. Zhang, "Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 11 753–11 762.

[36] H. Wang, P. Lu, H. Zhang, M. Yang, X. Bai, Y. Xu, M. He, Y. Wang, and W. Liu, "All you need is boundary: Toward arbitrary-shaped text spotting," in *Proc. AAAI*, 2020, pp. 12 160–12 167.

[37] E. Xie, W. Wang, M. Ding, R. Zhang, and P. Luo, "Polarmask++: Enhanced polar representation for single-shot instance segmentation and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[38] Z. Tian, M. Shu, P. Lyu, R. Li, C. Zhou, X. Shen, and J. Jia, "Learning shape-aware embedding for scene text detection," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 4234–4243.

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

[40] W. Wang, E. Xie, X. Li, X. Liu, D. Liang, Y. Zhibo, T. Lu, and C. Shen, "Pan++: Towards efficient and accurate end-to-end spotting of arbitrarily-shaped text," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[41] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, and S. Shao, "Shape robust text detection with progressive scale expansion network," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 9336–9345.

[42] R. Vatti, "A generic solution to polygon clipping," *Commun. ACM*, vol. 35, no. 7, pp. 56–63, 1992.

[43] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[44] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015, pp. 448–456.

[45] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. AISTATS*, 2011, pp. 315–323.

[46] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[47] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. Int. Conf. 3D Vis.*, pp. 565–571.

[48] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 1083–1090.

[49] 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, 2014.

[50] C. K. Ch'ng and C. S. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in *Proc. ICDAR*, vol. 1, 2017, pp. 935–942.

[51] Y. Liu, L. Jin, S. Zhang, and S. Zhang, "Detecting curve text in the wild: New dataset and new solution," *arXiv preprint arXiv:1712.02170*, 2017.

[52] D. Karatzas, L. Gomez, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. Chandrasekhar, and S. Lu, "Icdar 2015 competition on robust reading," in *Proc. ICDAR*, 2015, pp. 1156–1160.

[53] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 2315–2324.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.

[55] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *Proc. Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 248–255.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[58] W. Feng, W. He, F. Yin, X.-Y. Zhang, and C.-L. Liu, "Textdragon: An end-to-end framework for arbitrary shaped text spotting," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 9076–9085.