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 dependence 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 distinction 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, zoomed-out view module (ZOM) is introduced to provide coarse-grained optimization objectives for coarse layers to avoid feature defocusing. Meanwhile, zoomed-in view 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—Detail loss, false-positive samples, feature defocusing, text detection, zoom strategy.

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], [2], [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], [6] and comprehensive performance prior methods [7], [8], [9].

Manuscript received 6 June 2022; revised 25 April 2023; accepted 21 June 2023. Date of publication 4 July 2023; date of current version 30 October 2024. This work was supported in part by the National Natural Science Foundation of China under Grant U2112041, Grant 61825603, and Grant 62106182; and in part by the National Key Research and Development Program of China under Grant 2020YFB2103902. (Corresponding author: Qi Wang.)

Chuang Yang is with the School of Computer Science and the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi’an 710072, China (e-mail: cyang113@mail.nwpu.edu.cn).

Mulin Chen, Yuan Yuan, and Qi Wang are with the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi’an 710072, China (e-mail: chenmulin@mail.nwpu.edu.cn; y.yuan.ieee@gmail.com; crabwq@gmail.com).

Digital Object Identifier 10.1109/TNNLS.2023.3289327

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 dependence for high-performance computing units, and slow inference. Furthermore, related works [10], [11], [12] show 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 [8], [9] 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 multilevel feature maps and rebuild text contours by simple postprocessing, which simplifies the frameworks and improves the detection speed effectively. However, three disadvantages exist in these methods, which limit the improvement of detection accuracy.

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 (can be referred in the methodology section), 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 leads to the insufficient model’s ability to distinguish the foreground and the background.

Considering the limitations above, how to overcome those 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 article, 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 zoomed-out view 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 distinction of shrink-masks from the corresponding margins, a zoomed-in view module (ZIM) simulates the zoom-in process of the camera to utilize fine features to enhance ZTD’s ability to recognize the margins. Moreover, considering there are strong sequential feature differences between shrink-masks and noise regions, a sequential-visual discriminator (SVD) is designed to combine sequential and visual features to help suppress false-positive samples. The main contributions of this article are as follows:

1) Zoom strategy-based ZOM and 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 distinction of shrink-masks from the margins.

2) An 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 postprocessing is constructed. It achieves 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 this article 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 this article.
which could effectively fit arbitrary-shaped text instances and overcome the text adhesion problem. Although these methods enjoy high detection speed, pixelwise extension-based postprocessing is relatively time-consuming. Considering this problem, Liao et al. [9] designed an objectwise extension strategy-based text detection framework, which only needed to predict shrink-masks by one segmentation header. Importantly, the objectwise strategy saved much computational cost compared with [8].

III. METHODOLOGY

In this section, we introduce the overall structure of the proposed ZTD first. Then, the details of ZOM, ZIM, and 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 Fig. 2, which is composed of backbone, ZOM, ZIM, SVD, prediction header, and contour extension process. For backbone, it is used for the generation of multilevel feature maps $f_1$, $f_2$, $f_3$, and $f_4$ corresponding to the image size of $(1/4)$, $(1/8)$, $(1/16)$, and $(1/32)$, respectively. To avoid the phenomena of feature defocusing and detail loss that exist in current methods [8], [9], 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 in the training process, which helps 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 $f_4$ first. To save computational cost, we use two $3 \times 3$ 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 pointwise addition and experimental results in [42] 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 the concatenated feature channels to half of the original to generate the feature map $F_1$.

In the training process, a $3 \times 3$ convolution is used for $F_1$ smoothing, and a $1 \times 1$ 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 [41] and the region in shrunk contour is treated distinguished 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 [41].
as shrink-mask. The ground-truth of the coarse shrink-mask is obtained by resizing shrink-mask to (1/16) of image size. Particularly, except the final 1 × 1 convolution, each convolutional layer is followed by a BatchNorm layer [43] and rectified linear unit (ReLU) [44].

C. Zoomed-In View 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 map is treated as an input for the following two branches. For the first one, the feature map is downsampled by a $3 \times 3$ 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 map by transposed convolution as (1/4) size of the image for pixelwise segmentation. In the end, considering the image details are lost with the increase of network layers, the above concatenated feature map is skip-connected with the upsampled feature map $F_2$.

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; and 5) conducting “ors” operation between the processed text mask and the result generated by step 3. The mentioned “ors” operation is defined as

$$p_{i,j}^a \text{ ors } p_{i,j}^b = \text{ ignorance}$$

$$p_{i,j}^a \text{ ors } p_{i,j}^b = 1, \quad \{p_{i,j}^a = 1 \text{ or } p_{i,j}^b = 1\}$$

$$p_{i,j}^a \neq \text{ ignorance and } p_{i,j}^b \neq \text{ ignorance}$$

(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.

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 preprocessing and discriminator. The preprocessing takes (1/4) GT shrink-mask and $F_3$ (shown in Fig. 2) as input. It first executes multiplication on them to suppress false-positive samples. Particularly, SVD is a training-only module, which brings no extra computational cost 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

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

(3)

where the parameters $\alpha$, $\beta$, $\gamma$, and $\eta$ balance the importance of different loss functions. They are set to 1, 0.25, 0.25, and 0.25 in the following experiments, respectively.

1) **Optimization of Shrink-Mask Prediction Header:** Dice loss [46] 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

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

(4)

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

2) **Optimization of Zoomed-In View 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 text. 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 $\mathcal{L}_{zi}$ can be expressed as

$$
\mathcal{L}_{zi} = 1 - \frac{2 \times |ZI_p \cap ZI_g| + \varepsilon}{|ZI_p| + |ZI_g| + \varepsilon}
$$

(5)

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

3) **Optimization of Zoomed-Out View Module:** The label of this module is shrink-mask of $(1/16)$ 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 mask and ground-truth

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

(6)

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

4) **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 binary cross
entropy (BCE) loss to measure the loss $L_{svd}$ of this module

$$L_{svd} = -(S_p \times \log(S_p) + (1 - S_p) \times \log(1 - S_p))$$  \hspace{1cm} (7)

where $S_p$ is the probability whether the region is shrink-mask and $S_g$ 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** [47] is a dataset consisting of line-level Chinese and English text instances. There are 300 training images and 200 testing images, respectively. The same as previous works, we introduce 400 extra images from HUST-TR400 [48] as training data.

**Total-Text** [49] 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** [50] has 1000 training images and 500 testing images. Different from Total-Text, this dataset mainly consists of line-level arbitrary-shaped text instances.

**ICDAR2015** [51] 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. ICDAR2015 utilizes 1000 images to train the model and 500 images to evaluate the model 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 similar geometry characteristics. To ensure a fair comparison
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 ZTD IS PRETRAINED ON SYNTHTEXT [52]

| # | 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 (0.9%) | 64.1 |
| 3 | baseline+ | ✓   | ✓   |     |     | 90.7 | 80.3 | 85.2 (1.3%) | 59.2 |
| 4 | baseline+ | ✓   | ✓   | ✓   |     | 92.2 | 80.9 | 86.2 (1.0%) | 59.2 |
| 5 | baseline+ | ✓   | ✓   | ✓   | ✓   | 91.6 | 82.4 | 86.8 (0.6%) | 59.2 |

![Fig. 8](image-url)

**Fig. 8.** Visualization of the predicted shrink-masks of ZTD with different settings. The red circled regions show some false-positive samples are suppressed and hard-positive samples are recognized with the help of the designed ZIM, ZOM, and SVD. (a) RGB image. (b) Baseline. (c) +ZIM. (d) +ZOM. (e) +SVD. (f) +Ext. (g) GT shrink-mask.

In the initialization stage, we resize the short sides of original images to specific sizes to generate resized text instances. It can be found that the resized training and testing samples enjoy similar geometry characteristic distributions.

**B. Implementation Details**

The overall architecture of our method is shown in Fig. 2, where the feature maps \((f_1, f_2, f_3, \text{ and } f_4)\) behind input are generated by different stages \((\text{stage1, stage2, stage3, and stage4, respectively})\) of ResNet-18 [53].

In the data preprocessing 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 \((-10^\circ, 10^\circ)\); and 4) random cropping and padding.

In the initializing stage, the backbone of ZTD is pretrained on ImageNet [54] and the rest of the layers are initialized by the strategy proposed in [55]. In the training process, the Adam 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 pretrained 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 Figs. 2 and 4–6 are abandoned, which is helpful to facilitate detection speed.

**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 subloss of \(L\) and the importances of different RNN units of SVD, respectively. The details of experimental results are described in next.

1) Effectiveness of Zoomed-In View 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 (see 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 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.

2) Effectiveness of Zoomed-Out View 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.

3) 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 pretrain our model on SynthText in this article to keep a fair comparison environment with existing methods. As shown in Fig. 8(f), pretraining our model on SynthText improve the accuracy of predicted shrink-masks and brings 0.6% F-measure (see Table I #5). Furthermore, we can see from Fig. 9(a), the pretrained model converges more faster than others.

4) 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.

5) Importances of Different Sublosses: As described in Section III-E, the optimization function L is composed of Lsm, Lz2, Lzo, and Lsvd, 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 subloss. 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 importances of different sublosses intuitively.
TABLE II

| Methods          | P  | R  | F  | FPS |
|------------------|----|----|----|-----|
| **Accuracy Prior** |    |    |    |     |
| PixelLink [29] (AAAI 2018) | 83.0 | 73.2 | 77.8 |   |
| RBD [24] (CVPR 2018) | 87.0 | 73.0 | 79.0 | 10 |
| CRAFT [32] (CVPR 2018) | 88.2 | 78.2 | 82.9 | 8.6 |
| SAE [37] (CVPR 2019) | 84.2 | 81.7 | 82.9 |   |
| TextField [38] (TIP 2019) | 87.4 | 75.9 | 81.3 |   |
| OPMP [6] (TMM 2020) | 86.0 | 83.4 | 84.7 | 1.6 |
| SAVTD [5] (CVPR 2021) | 89.2 | 81.5 | 85.2 |   |
| **GV [25] (TPAMI 2021)** | 88.8 | 84.3 | 86.5 | 15.0 |

**Comprehensive Performance Prior**

| Methods   | P  | R  | F  | FPS |
|-----------|----|----|----|-----|
| DB [9] (AAAI 2020) | 90.4 | 76.3 | 82.8 | 62.0 |
| PAN [8] (ICCV 2019) | 84.4 | 83.8 | 84.7 | 30.2 |
| PAN++ [39] (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.

1) Evaluation on MSRA-TD500: We evaluate the performance of ZTD for detecting multilanguage 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 gliding vertex (GV) [25], 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 postprocessing, our method runs four times faster than it. Furthermore, the comprehensive performance of ZTD-736 outperforms pixel aggregation network (PAN) [8], PAN++ [29] a lot. Though differentiable binarization (DB) [9] achieves 62.0 frames/s in detection speed, ZTD-512 is 35.4 frames/s faster than it. Some qualitative results are shown in Fig. 12(a). The above results demonstrate the effectiveness of ZTD for detecting long text instances, even they are multilingual.

2) 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 [12], ContourNet [34], and deep relational reasoning graph (DRRG) [30] achieve 85.2%, 85.4%, and 85.7% in F-measure, respectively. For comprehensive performance prior approaches, PAN [8] and PAN++ [29] enjoy comparable detection accuracy with accuracy prior methods. DB [9] performs better in detection speed. Compared with the above methods, ZTD can achieve 86.0% in F-measure and 75.2 frames/s. 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 problems. We display some detection results in Fig. 12(b). It is found that adhesive texts are separated successfully.

3) 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 frames/s and is faster than other methods at least by 21.9 frames/s. The outstanding detection speed benefits from the efficient text representation method and lightweight CNN model. Specifically, PAN [8] and PAN++ [29] reconstruct text contours through pixelwise extension strategy. Unlike these comprehensive performance prior methods, ZTD adopts an objectwise 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 detection accuracy of ZTD is not as well as some methods (such as DRRG [30] and exploiting visual relationships for arbitrary-shaped scene text detection with graph convolutional networks (ReLaText) [11]), the proposed approach has at least seven 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.

4) Evaluation on ICDAR2015: To verify the robustness of ZTD to detect multioriented text instances from 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 frames/s, which outperforms DB [9] and PAN [8] in both detection accuracy and speed. Moreover, ZTD can run two times faster than PAN++ [29] and can achieve 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 multiorientations 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 [29], RRD [24], and TextField [38]) that is
trained on MSRA-TD500, which shows the effectiveness of 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 objectwise 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 structure and objectwise contour extension process bring significant improvements to inference speed, and the results demonstrate this conclusion.

G. Failure 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 blur and dim light environment are the failure reasons. For line-level (CTW1500) and word-level (Total-Text) datasets, half detection, and overdetection are the current main problems, respectively. Except for the above problems, low color contrast brings challenges for scene text detection tasks. As shown in the MSRA-TD500 sample of Fig. 13, our model is hard to recognize the text instance accurately because of the highly similar color between the foreground and the background. The above issues make there is still much room to improve the proposed method.

V. Conclusion

In this article, we propose an efficient text detector inspired by the zoom process of the camera, termed as 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. Ablation studies show the effectiveness of ZOM, ZIM, and SVD. Comparisons on the multiple benchmarks verify the superior performance of ZTD in both detection accuracy and speed, which shows the great potential for a wide range of applications.

TABLE VII

| Datasets   | Image size (pixel) | 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  | 1.4  | 48.3 | 83.0 |

Fig. 13. Challenging samples from ICDAR2015, CTW1500, Total-Text, and MSRA-TD500 datasets. The green bounding boxes are the detection results from our method. The yellow ones are labels.

References

[1] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 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. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.
[3] Z. Tian, C. Shen, H. Chen, and T. He, “FCOS: Fully convolutional one-stage object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 9626–9635.
[4] S. Zhang, X. Zhu, J. Hou, C. Yang, and X. Yin, “Kernel proposal network for arbitrary shape text detection,” IEEE Trans. Neural Netw. Learn. Syst., early access, Mar. 10, 2022, doi: 10.1109/TNNLS.2022.3152596.
[5] W. Feng, F. Yin, X. Zhang, and C. Liu, “Semantic-aware video text detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 1695–1705.

[6] 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, 2021.

[7] X. Zhou et al., “EAST: An efficient and accurate scene text detector,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2692–2701.

[8] W. Wang et al., “Efficient and accurate arbitrary-shaped text detection with pixel aggregation network,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (ICCV), Oct. 2019, pp. 8439–8448.

[9] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, “Real-time scene text detection with differentiable binarization,” in Proc. AAAI, 2020, pp. 11474–11481.

[10] Y. Zhu, J. Chen, L. Liang, Z. Kuang, L. Jin, and W. Zhang, “Fourier contour embedding for arbitrary-shaped text detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 3122–3130.

[11] C. Ma, L. Sun, Z. Zhong, and Q. Huo, “ReLaText: Exploiting visual relationships for arbitrary-shaped scene text detection with graph convolutional networks,” Pattern Recognit., vol. 111, Mar. 2021, Art. no. 107684.

[12] S. Liu, 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, Feb. 2021.

[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 Inform. Process. Syst., 2015, pp. 91–99.

[14] Y. Liu and L. Jin, “Deep matching prior network: Toward tighter multi-oriented text detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3454–3461.

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

[16] C. Zhang et al., “Look more than once: An accurate detector for text of arbitrary shapes,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 10544–10553.

[17] B. Shi, X. Bai, and S. Belongie, “Detecting oriented text in natural images by linking segments,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3482–3490.

[18] S. Tian, S. Lu, and C. Li, “WeText: Scene text detection under weak supervision,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1501–1509.

[19] W. Liu et al., “SSD: Single shot MultiBox detector,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 21–37.

[20] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, “TextBoxes: A fast text detector with a single deep neural network,” 2016, arXiv:1611.06779.

[21] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, “Single shot text detector with regional attention,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3066–3074.

[22] 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.

[23] P. Wang, H. Li, and C. Shen, “Towards end-to-end text spotting in natural scenes,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 10, pp. 7266–7281, Oct. 2022.

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

[25] Y. Xu et al., “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, Apr. 2021.

[26] F. Wang, Y. Chea, F. Wu, and X. Li, “TextRay: Contour-based geometric modeling for arbitrary-shaped scene text detection,” in Proc. 28th ACM Int. Conf. Multimedia, 2020, pp. 111–119.

[27] L. Huang, Y. Yang, Y. Deng, and Y. Yu, “DenseBox: Unifying landmark localization with end to end object detection,” 2015, arXiv:1509.04874.

[28] X. Li, M. Chen, F. Nie, and Q. Wang, “A multi-view based parameter free framework for group detection,” in Proc. AAAI Conf. Artif. Intell., Feb. 2017, vol. 31, no. 1, pp. 4147–4153.

[29] D. Peng, H. Liu, X. Li, and D. Cai, “Pixellink: Detecting scene text via instance segmentation,” in Proc. AAAI, 2018, pp. 6773–6780.
Chuang Yang received the B.E. degree in automation and the M.E. degree in control engineering from the Civil Aviation University of China, Tianjin, China, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree with the School of Computer Science and the School of Artificial Intelligence, OPTics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi’an, China. His research interests include computer vision and machine learning.

Mulin Chen received the B.E. degree in software engineering and the Ph.D. degree in computer application technology from Northwestern Polytechnical University, Xi‘an, China, in 2014 and 2019, respectively. He is currently a Researcher with the School of Artificial Intelligence, OPTics and ElectroNics (iOPEN), Northwestern Polytechnical University. His current research interests include computer vision and machine learning.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, OPTics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi’an, China. His research interests include computer vision, pattern recognition, and remote sensing.

Yuan Yuan (Senior Member, IEEE) is currently a Full Professor with the School of Artificial Intelligence, OPTics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi’an, China. She has authored or coauthored over 150 articles, including about 100 in reputable journals, such as the IEEE TRANSACTIONS and Pattern Recognition, as well as the conference papers in IEEE Conference on Computer Vision and Pattern Recognition (CVPR) and British Machine Vision Conference (BMVC). Her current research interests include visual information processing and image/video content analysis.