TU Delft

Delft University of Technology

Accurate Scene Text Detection via Scale-Aware Data Augmentation and Shape Similarity Constraint

Dai, Pengwen; Li, Yang; Zhang, Hua; Li, Jingzhi; Cao, Xiaochun

DOI
10.1109/TMM.2021.3073575

Publication date
2021

Document Version
Accepted author manuscript

Published in
IEEE Transactions on Multimedia

Citation (APA)
Dai, P., Li, Y., Zhang, H., Li, J., & Cao, X. (2021). Accurate Scene Text Detection via Scale-Aware Data Augmentation and Shape Similarity Constraint. *IEEE Transactions on Multimedia*, *24*, 1883-1895. https://doi.org/10.1109/TMM.2021.3073575

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

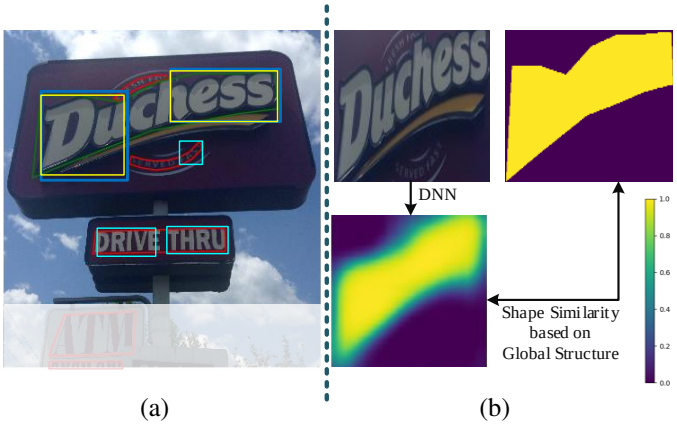
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3073575, IEEE Transactions on Multimedia

Accurate Scene Text Detection via Scale-Aware Data Augmentation and Shape Similarity Constraint

Pengwen Dai, Yang Li, Hua Zhang, Jingzhi Li, Xiaochun Cao, *Senior Member, IEEE*

***Abstract*—Scene text detection has attracted increasing concerns with the rapid development of deep neural networks in recent years. However, existing scene text detectors may overfit on the public datasets due to the limited training data, or generate inaccurate localization for arbitrary-shape scene texts. This paper presents an arbitrary-shape scene text detection method that can achieve better generalization ability and more accurate localization. We first propose a Scale-Aware Data Augmentation (SADA) technique to increase the diversity of training samples. SADA considers the scale variations and local visual variations of scene texts, which can effectively relieve the dilemma of limited training data. At the same time, SADA can enrich the training minibatch, which contributes to accelerating the training process. Furthermore, a Shape Similarity Constraint (SSC) technique is exploited to model the global shape structure of arbitrary-shape scene texts and backgrounds from the perspective of the loss function. SSC encourages the segmentation of text or non-text in the candidate boxes to be similar to the corresponding ground truth, which is helpful to localize more accurate boundaries for arbitrary-shape scene texts. Extensive experiments have demonstrated the effectiveness of the proposed techniques, and state-of-the-art performances are achieved over public arbitrary-shape scene text benchmarks (e.g., *CTW1500*, *Total-Text* and *ArT*).**

***Index Terms*—Scene text detection, arbitrary shape, text part, global context, data augmentation, accurate localization.**



(a) (b)

Fig. 1: (a) Scale-aware data augmentation. Each image crop only cares about some specific-scale scene texts (green) and ignores other texts (red). The positive (yellow) text parts and the negative (cyan) text parts will be considered in the training process. These text parts are generated based on the anchors (blue) and the scene texts. (b) Shape similarity constraint. The generated segmentation map is encouraged to be similar to the ground-truth mask. Best view in color.

I. Introduction

SCENE text reading plays a significant role in many practical applications, such as scene understanding [1], image retrieval [2], autonomous driving [3], etc. Scene text detection, as the prerequisite of the scene text reading system, has attracted increasing interests in the field of computer vision and multimedia. With the development of deep neural networks, many scene text detection approaches [4]–[9] are proposed. Although some impressive results have been achieved, there are still several challenges for developing a robust and accurate scene text detector.

The first challenge is the limited number of training data on the standard benchmarks, e.g., 229 training images in *ICDAR2013* [10], 1000 training images in *CTW1500* [11]. When training a deep neural network on these datasets, the model may be overfitting. To solve this problem, one simple strategy is to increase the number of training samples by using data augmentation, which includes scaling, rotating, cropping and flipping of training images. The other strategy is to employ the synthetic data [12] to pre-train the model and then fine-tune on the real-world training data. However, the traditional data augmentation strategies focus on the global variations of annotated texts, which may be failed for the local visual variation cases. Moreover, since the synthetic data [12] is usually large-scale (∼800,000 images), pre-training on this synthetic data is time-consuming.

The second challenge is to accurately localize the arbitrary-shape scene text. The layout of the text line or text word in the natural image would be arbitrary-shape, which is hard to be accurately localized by the horizontal or multi-oriented scene text detectors [4]–[7], [13]–[25]. To achieve accurate localization, the regression of multiple key points [11], [26] is introduced to fit the curved bounding boxes of scene texts,

Manuscript received March 9, 2020; revised July 2, 2020 and April 7, 2021; accepted April 11, 2021. This work was supported by the National Key R&D Program of China (Grant No. 2020YFB1406704), National Natural Science Foundation of China (No. 62025604, 61733007, 62072454, U1936208, U1736219), Key Program of the Chinese Academy of Sciences (No. QYZDB-SSW-JSC003), Beijing Natural Science Foundation (No. 4202084), Peng Cheng Laboratory Project of Guangdong Province PCL2018KP004. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. David Crandall. (Corresponding author: Xiaochun Cao.)

P. Dai, H. Zhang, J. Li, and X. Cao are with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100195, China; and also with the University of Chinese Academy of Sciences, Beijing 100049, China. X. Cao is also with the Peng Cheng Laboratory, Cyberspace Security Research Center, Shenzhen 518055, China. (e-mail: daipengwen@iie.ac.cn, zhanghua@iie.ac.cn, caoxiaochun@iie.ac.cn).

Y. Li is with the Algorithmics group at Delft University of Technology, 2628 XE Delft, The Netherlands. (e-mail: Y.Li-31@tudelft.nl)

1520-9210 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: TU Delft Library. Downloaded on May 06,2021 at 19:28:26 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3073575, IEEE Transactions on Multimedia

but their localization boundaries are sensitive to the prediction of each key point. Moreover, there are also some solutions [8], [9], [27]–[37] to extract the text regions by performing pixel-wise segmentation. However, the segmentation-based methods are difficult to separate the neighboring scene text instances and may generate the over-segmentation or under-segmentation, due to the local-aware cues and the no well-defined closed geometry boundary of the scene text.

To solve the above-mentioned challenges, two novel techniques are introduced. Specifically, we first design a novel **S**cale-**A**ware **D**ata **A**ugmentation (SADA) strategy for the task of arbitrary-shape scene text detection. It not only considers the extreme variety of scales and aspect ratios of scene texts but also regards scene text parts as new scene texts, which makes full use of the scale variation and local visual variation of scene texts. As illustrated in Fig. 1(a), SADA generates a fixed-scale image crop from the input images with different scales. In each image crop, only some specific-scale scene texts participate in the training process. Based on these specific-scale texts (green), their parts (yellow) are considered to increase the positive training samples. For the scene texts (red) that do not fall into a specific scale, their parts (cyan) are employed to remove ambiguous negative samples. These parts are aware of the scale of anchors (blue) and scene texts. Since the low-resolution image crop (512×512) enlarges the minibatch size and the text parts increase the diversity of training samples, SADA can lead our model to learn more text variations in each iteration, which will accelerate the convergence and promote the performance of the model.

Next, to accurately localize the boundaries of arbitrary-shape scene texts, we exploit a novel **S**hape **S**imilarity **C**onstraint (SSC) loss function. SSC encourages the segmentation map of text/non-text generated by the deep neural network (DNN), to be similar to the corresponding ground-truth mask, as shown in Fig. 1(b). The advantages of SSC are that it can capture the global context from the perspective of the loss function, which will not introduce extra network parameters and calculations in the inference stage. Moreover, SSC models the global shape structures, which helps to learn discriminative and robust feature representations.

The contributions of our work are summarized as follows:

i) We propose a scale-aware data augmentation technique, which accelerates the training and improves the performance.

ii) A shape similarity constraint is exploited to capture global shape structures, which is helpful to achieve more accurate localization.

iii) Our model can detect arbitrary-shape scene texts, and has achieved state-of-the-art performances on the public arbitrary-shape scene text benchmarks.

The rest of the paper is organized as follows. We first introduce the related work in Section II. Then in Section III, the proposed method is presented in detail. In Section IV, we conduct numerous experiments and describe experimental results. Section V finally concludes the paper.

II. Related work

In the era of deep learning, most scene text detection methods can be found in the recent survey [38]. These studies are roughly divided into two mainstreams: regression-based methods and segmentation-based methods.

For the regression-based methods, researchers [4], [13] usually take the scene text as a special object and inherit the frameworks of general object detection to detect the horizontal scene text. Researchers [5], [7], [14]–[19], [23]–[25] also propose multi-oriented scene text detection techniques, such as rotating anchors [5], rotating the convolution filters [16], learning the affine transformation of the bounding box [19] and regressing the angles or corner points of inclined boxes [18], [22]. When detecting the arbitrary-shape scene text, some researchers further regress the locations of key points in the bounding box [11], [26]. Besides, some methods [6], [20], [21] first regress scene text parts, and then aggregate the detected parts into the horizontal, multi-oriented or arbitrary-shape text instances.

For the segmentation-based methods, some researchers directly segment the text regions from the entire input image. Instead of only performing semantic segmentation for each pixel, more excellent approaches are proposed to learn more attributes, such as learning the link relationship among pixels [39], predicting the text border [24], learning the geometry attributes [7], [8], [27]–[32] of each pixel, constructing text instance with the progressive scale expansion [9], pulling pixels of the same text and pushing pixels of different text instances [40], [41], and so on. Besides, some methods perform the segmentation only on the bounding box. Inheriting the framework of instance-aware semantic segmentation, some efficient methods are proposed [33]–[36] for detecting scene text. To alleviate the labor of designing anchors, Tian *et al.* [37] employ an anchor-free network to generate the candidate boxes, and propose an iterative refinement module to obtain a more accurate localization of the bounding box.

In addition, data augmentation is a routine operation in the deep learning to avoid over-fitting. Besides the traditional data augmentation strategies implemented by rotating, cropping, translating, scaling and flipping images, some special data augmentation methods [42], [43] are proposed for various computer vision tasks. In the field of scene text detection, existing methods usually utilize the synthetic data to augment the training samples. Some techniques [12], [44]–[46] are proposed to synthesize scene text images by embedding various texts into natural images automatically. However, the synthetic data usually exists gaps with the real data, which can not ensure the consistency of data distribution. So far, only a few works [24], [47] pay attention to generating data from existing scene text images. Zhan *et al.* [47] generate scene text images by transforming the source domain to the target domain in both appearance and geometry spaces. They exploit a kind of generative adversarial network to achieve the cross-domain shifts, so the generated data can still be regarded as the synthetic data in essence. To generate real training data, Xue *et al.* [24] design a bootstrapping strategy by randomly sampling text parts and repainting the unsampled ones.

In this work, we propose a method to detect arbitrary-shape scene texts. It is a segmentation-based method that performs the segmentation on the bounding boxes. The proposed scale-aware data augmentation (SADA) mechanism considers the

1520-9210 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: TU Delft Library. Downloaded on May 06,2021 at 19:28:26 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3073575, IEEE Transactions on Multimedia

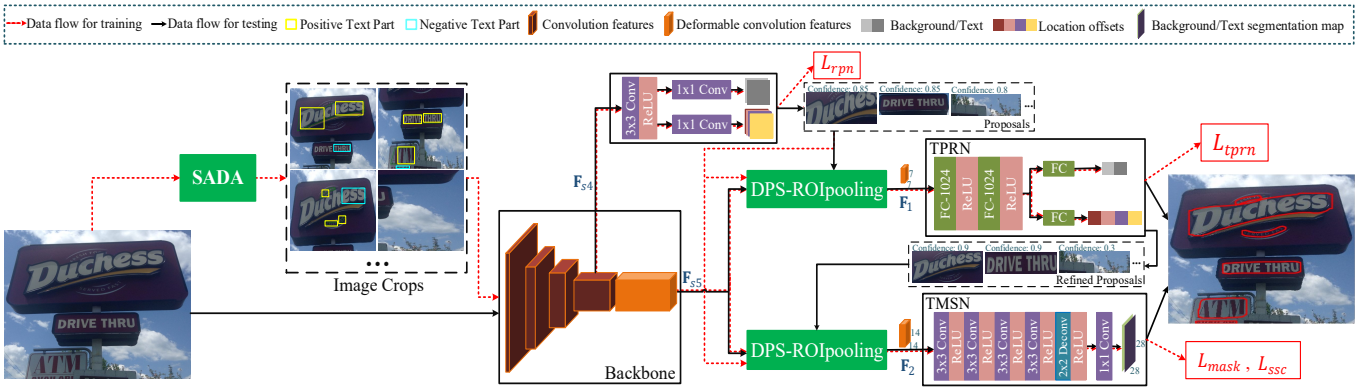


Fig. 2: Overview of our architecture. In the training stage, the scale-aware data augmentation (SADA) is first employed to generate various image crops and text parts (yellow and cyan denote the positive and negative text parts). Then, for any image crop, the backbone network is utilized to extract features. After that, according to the extracted features, the region proposal network (RPN) generates the proposals, which will be projected into fixed-scale features by the deformable position-sensitive region-of-interest pooling (DPS-ROIpooling). Finally, based on the different fixed-scales features, the text proposal refinement network (TPRN) is adopted to generate the probabilities of classes and the location offsets, while the text mask segmentation network (TMSN) is employed to generate the text mask. Furthermore, the shape similarity constraint serves as a loss function L_{ssc} to learn the global shape structure. In the inference, the input image is directly fed into the backbone network, and the refined proposals generated by TRPN will be fed into the TMSN to generate the segmentation mask of arbitrary-shape scene texts. Best view in color.

specific scales and scene text parts. It can be seen as an image cropping strategy but involves innovative designs based on the characteristics of scene texts. Although the researchers in [24] have utilized the text parts, our differences are that: i) Xue *et al.* [24] generate image crops via a simple multi-scale cropping strategy, while our SADA generates positive and negative image crops based on the scale and local visual variations of scene texts. ii) In the training process, Xue *et al.* [24] employ all various-scale scene texts in each image crop, while our SADA only cares about the specific-scale scene texts. iii) Xue *et al.* [24] randomly sample text parts and then repaint the remained ones for scene texts, while our SADA generates text parts based on the overlaps between the scene texts and anchors. iv) Xue *et al.* [24] mainly utilize the augmentation scheme to improve the consistency of the predicted text feature map. However, our SADA is aware of the scene texts and anchors, which is used to increase the number of positive training samples and decrease the confusion of negative training samples. Furthermore, we also propose the novel shape similarity constraint to detect more accurate boundaries for arbitrary-shape scene text instances.

III. Methodology

A. Overview of Architecture

The architecture of our method is illustrated in Fig. 2. Specifically, given an input image $\mathbf{I}$ with the height H and the width W, we extract the representative feature maps by using the backbone network incorporated with deformable convolutions. Then, the feature $\mathbf{F}_{s4} \in \mathbb{R}^{\frac{H}{S} \times \frac{W}{S} \times D_1}$ derived from the stage-4 of the backbone network is fed into the region proposal network (RPN) [48] to generate proposals. Next, the corresponding fixed-scale feature is extracted from the stage-5 feature $\mathbf{F}_{s5} \in \mathbb{R}^{\frac{H}{S} \times \frac{W}{S} \times D_2}$ of the backbone network by using the deformable position-sensitive region-of-interest pooling (DPS-ROIpooling) [49]. Note that the learnable parameters in the different DPS-ROIpooling modules are not shared. Finally, the fixed-scale feature $\mathbf{F}_1 \in \mathbb{R}^{\rho_1 \times \rho_1 \times D_3}$ is fed into the text proposal refinement network (TPRN) to generate the probabilities of text/non-text and location offsets of text, while the fixed-scale feature $\mathbf{F}_2 \in \mathbb{R}^{\rho_2 \times \rho_2 \times D_3}$ is fed into the text mask segmentation network (TMSN) to predict the segmentation mask. Note that the TPRN and TMSN have the same architectures with the RCNN branch and the mask branch in [50], respectively. Besides, in the step of training, we utilize the scale-aware data augmentation technique to increase the diversity of training samples, and we also employ the shape similarity constraint to learn the global shape structures for text and non-text.

B. Scale-Aware Data Augmentation

The scales and aspect ratios of scene texts are extremely various, and the scene text parts can still be regarded as new texts due to lack of well-defined closed boundaries of texts. Based on these characteristics of scene texts, we propose the scale-aware data augmentation (SADA) to increase the diversity of training samples and facilitate the training process. The procedure of SADA is shown in Fig. 3, which involves two main steps: image crop generation and text part generation.

1) Image Crop Generation: We first construct the image pyramid to generate different scales of inputs, denoted as $\{S_i \mid i{=}1, 2, 3\}$. For each scale S_i, a K×K sliding window is used to extract the image crops $\mathcal{C}_i$. Furthermore, we design a short edge range $R_i = [e_i^{min}, e_i^{max}]$ of the bounding box to determine the ground-truth boxes of participating in the training process. These concerned ground-truth boxes are denoted as $\mathcal{G}_i$ for each scale. Finally, the crops are selected

1520-9210 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: TU Delft Library. Downloaded on May 06,2021 at 19:28:26 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3073575, IEEE Transactions on Multimedia

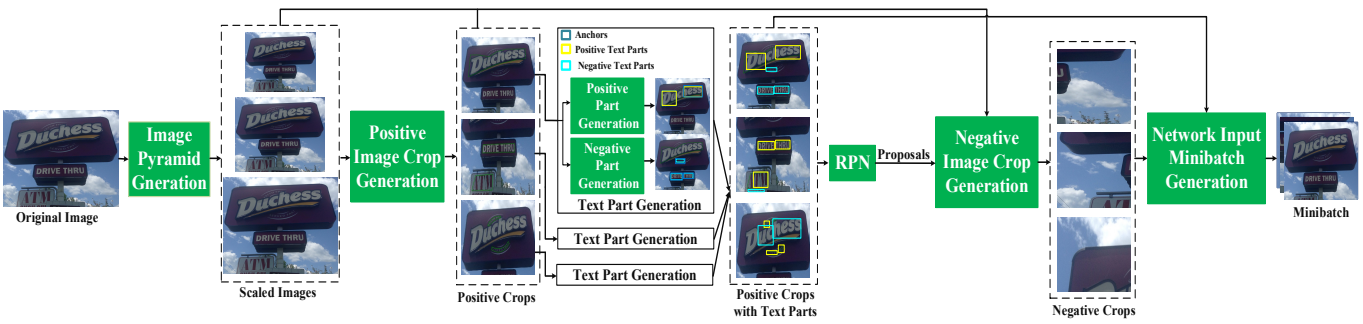


Fig. 3: Procedure of the scale-aware data augmentation. The original input image is first scaled to different-scale images. Next, the scaled images are used to generate positive image crops. Subsequently, the text part generation module is employed to generate positive and negative text parts. After that, the positive crops including text parts are utilized to train a region proposal network (RPN) for generating proposals. Then, the negative image crop generation module takes the scaled images, positive crops and proposals as inputs to obtain negative crops. Finally, we randomly sample from positive crops with text parts and negative crops to create the network input minibatch. Best view in color.

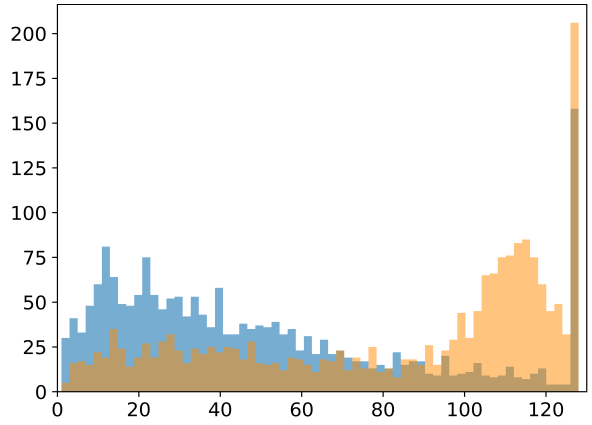


Fig. 4: Distribution against the number of positive training samples in RPN. When text parts are considered, they can generate more number of positive training samples (orange distribution) than that without considering text parts (blue distribution). The x-axis denotes the number of positive training samples while the y-axis is the number of image crops.

Algorithm 1 Positive Text Part Generation

Input: The set of total anchors $\{a_i\}$. The concerned ground-truth polygons $\{g_k\}$ with the widths $\{\varpi_k\}$ of the corresponding bounding boxes. The thresholds τ_o and τ_w.
Output: The positive anchor set $\mathcal{A}$. The text part set $\mathcal{T}$.

Set $\mathcal{A} \leftarrow \varnothing$, $\mathcal{T} \leftarrow \varnothing$.
for $\forall\ g_k$ **do**
  Generate the mask m_k based on g_k.
  Obtain the contour points $\{\mathbf{X}_k, \mathbf{Y}_k\}$ of m_k.
  for $\forall\ a_i$ **do**
    Obtain $(x_{min}, y_{min}, x_{max}, y_{max})$ of a_i.
    Set $\mathcal{S} \leftarrow \varnothing$.
    for $\forall\ (x', y')$ in $\{\mathbf{X}_k, \mathbf{Y}_k\}$ **do**
      if $x_{min} \leqslant x' \leqslant x_{max}$ **then**
        Append (x', y') to $\mathcal{S}$.
      end if
    end for
    Represent $(x_{min}, y_{min}, x_{max}, y_{max})$ of $\mathcal{S}$ as t.
    Compute the IOU overlap o between a_i and t.
    if $o > \tau_o$ and $x^t_{max} - x^t_{min} > \tau_w * \varpi_k$ **then**
      Set $\mathcal{A}_i \leftarrow a_i$, $\mathcal{T}_i \leftarrow t$.
    end if
  end for
end for

as the positive crops $\mathcal{C}_i^{pos}$ [51] if they cover the maximum number of $\mathcal{G}_i$. Note that a ground-truth bounding box may be covered by multiple crops with different scales due to the overlapping intervals in the consecutive R_i. For the unselected crops, some crops contain text-like regions. When these crops as negative crops participate in the training, they are helpful to improve the performance. To select these negative crops, we perform a negative crop mining. Specifically, we first employ the positive crops that contain text parts to train a region proposal network (RPN) [48] for generating proposals. Then, we remove the proposals that are completely covered by $\mathcal{C}_i^{pos}$ for each scale. After that, the crops are greedily selected when they are covered by the proposals in the range R_i. These selected crops are denoted as negative crops $\mathcal{C}_i^{neg}$. In the training, these positive crops and negative crops are randomly selected to create the network input minibatch. In experiments, the minibatch size is set to 10. The proportion of positive and negative crops is 4:1.

2) Text Part Generation: For most of the positive image crops, they usually contain a few concerned ground-truth boxes, which will result in generating limited positive training samples. To increase the diversity of positive training samples, we exploit the local visual variations of scene texts. Specifically, as illustrated in Fig. 3, for each concerned ground-truth poly (green), when the contour points are in the range of an anchor (blue) along the x-axis, these contour points are enclosed by a bounding rectangle (yellow), denoted as the part-text candidate. The anchor will be regarded as the positive training sample and the part-text candidate becomes the positive text part, when they satisfy two conditions: i) The IOU between the anchor and the part-text candidate is greater than a threshold τ_o. ii) The ratio between the width of the part-text candidate and the width of the bounding box of the ground-truth polygon is greater than a threshold τ_w. In experiments, we set $\tau_o = 0.7$ and $\tau_w = 1/3$. The whole

1520-9210 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: TU Delft Library. Downloaded on May 06,2021 at 19:28:26 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3073575, IEEE Transactions on Multimedia

process is summarized in Alg. 1. We also generate the training samples based on the ground-truth bounding boxes in the same way as [48]. When the number of positive training samples is not sufficient, we will employ Alg. 1 to increase the positive training samples, as shown in Fig. 4. Besides, to relieve the confusion of the negative training samples, we utilize the unconcerned ground-truth polygons (not labeled in green) instead of the concerned ground-truth polygons, to conduct the Alg. 1 for generating the negative text parts (cyan) and the negative training samples (blue). These negative training samples are ignored in the training.

C. Shape Similarity Constraint

For most scene text detection methods that perform segmentation on the proposals, they employ the pixel-wise binary cross-entropy loss, which is local-aware and lacks the global context. To better learn discriminative representations of each position in the segmentation map, we introduce the shape similarity constraint (SSC) to capture the global context of the predicted map from the perspective of the loss function. Inspired by SSIM [52] that is originally used to assert the image quality, our SSC is designed for arbitrary-shape scene text detection. It not only constrains the shape similarity of scene texts, but also constructs the shape similarity loss for the backgrounds. Specifically, SSC can be formulated as,

$$L_{ssc} = \frac{1}{C}\sum_{c=1}^{C} 1 - \phi\left(\frac{(2\cdot \mathbf{U}_p^c \odot \mathbf{U}_g^c + \varepsilon_1)(2\cdot \mathbf{Q}_{pg}^c + \varepsilon_2)}{(\varphi(\mathbf{U}_p^c, \mathbf{U}_g^c) + \varepsilon_1)(\mathbf{Q}_p^c + \mathbf{Q}_g^c + \varepsilon_2)}\right), \tag{1}$$

where C is the number of classes (text/non-text). ε_1 and ε_2 are factors to stabilize the division with weak denominator, which are fixed to 0.01^2 and 0.03^2 in experiments, respectively. ϕ denotes the average operation. φ represents the operation formulated as,

$$\varphi(\mathbf{U}_p^c, \mathbf{U}_g^c) = \mathbf{U}_p^c \odot \mathbf{U}_p^c + \mathbf{U}_g^c \odot \mathbf{U}_g^c, \tag{2}$$

where $\odot$ represents the element-wise multiplication. In addition, $\mathbf{U}_p^c, \mathbf{U}_g^c \in \mathbb{R}^{d\times d}$ denote the weighted mean maps for the predicted and ground-truth map of the c-th class. $\mathbf{Q}_p^c, \mathbf{Q}_g^c \in \mathbb{R}^{d\times d}$ are the weighted variance maps for the predicted and ground-truth map of the c-th class. $\mathbf{Q}_{pg}^c \in \mathbb{R}^{d\times d}$ is the weighted covariance map between the predicted map and the ground-truth map of the c-th class. They are calculated as,

$$\mathbf{U}_p^c = \mathbf{P}^c * \boldsymbol{\omega}, \;\; \mathbf{U}_g^c = \mathbf{G}^c * \boldsymbol{\omega}, \tag{3}$$

$$\mathbf{Q}_{pg}^c = (\mathbf{P}^c \odot \mathbf{G}^c) * \boldsymbol{\omega} - \mathbf{U}_p^c \odot \mathbf{U}_g^c, \tag{4}$$

$$\mathbf{Q}_p^c = (\mathbf{P}^c \odot \mathbf{P}^c) * \boldsymbol{\omega} - \mathbf{U}_p^c \odot \mathbf{U}_p^c, \tag{5}$$

$$\mathbf{Q}_g^c = (\mathbf{G}^c \odot \mathbf{G}^c) * \boldsymbol{\omega} - \mathbf{U}_g^c \odot \mathbf{U}_g^c, \tag{6}$$

where $*$ denotes the operation of correlation. $\boldsymbol{\omega} \in \mathbb{R}^{z\times z}$ is the Gaussian weighting filter with the size z, which is generated as the same with [52]. $\mathbf{P}^c \in \mathbb{R}^{d\times d}$ is the predicted map of the c-th class, whose element has the value of [0, 1]. $\mathbf{G}^c \in \mathbb{R}^{d\times d}$ is the ground-truth map of the c-th class, whose element has the value of 0 or 1.

D. Training and Inference

Our model is trained in an end-to-end manner. The total loss function is expressed as,

$$L = L_{rpn} + \lambda_1 L_{tprn} + \lambda_2 L_{tmsn}, \tag{7}$$

where λ_1 and λ_2 are the balance factors of the loss function, and both of them are set to 1 in experiments. L_{rpn} is the loss function of the region proposal network [48]. L_{tprn} denotes the loss function of the text proposal refinement network, which is similar to the RCNN loss [48]. The difference between them is that the former only involves two classes (text/non-text) while the latter includes multiple classes. L_{tmsn} is the loss function of the text mask segmentation network, which is calculated as,

$$L_{tmsn} = L_{mask} + \lambda_3 L_{ssc}, \tag{8}$$

where L_{mask} is the pixel-wise mask segmentation loss [50]. L_{ssc} is the loss function of the shape similarity constraint as illustrated in the above subsection. λ_3 is the balance factor between L_{mask} and L_{ssc}. In experiment, λ_3 is set to 1.

In the inference, the original input image is resized into the images with three scales $\{S_i \mid i\text{=}1,\ 2,\ 3\}$. For each scale, the detections are obtained from the predictions of TPRN and TMSN as [50]. Then all the detections are aggregated and filtered by the confidence threshold τ_c. After that, the mask-level NMS [53] is utilized to filter the overlapped detections. Finally, we utilize the marching square algorithm [54] to extract the contour for each detected binary mask. Considering the shape of the extracted contour, we define it as the polygonal bounding box. For the binary mask that may not have a completely connected component, it will generate multiple polygonal bounding boxes. Thus, we only keep the polygonal bounding box with the maximum area. In the post-processing, the mask-level NMS can effectively remove the detected text parts contained in the detected text instance, which will improve the precision of detection. The mask-level NMS calculates the overlaps as the following,

$$O = max(A_o/A_a, A_o/A_b), \tag{9}$$

where A_a and A_b are the areas of the detected masks a and b. A_o is the area of intersection between a and b. When O is greater than the threshold τ_n, the detections with lower confidences will be filtered.

IV. Experiments

In this section, we conduct experiments to demonstrate the effectiveness and superiority of our proposed method. We also analyze the limitations and the runtime of our model.

A. Datasets and Evaluation Protocols

CTW1500 [11] is an arbitrary-shape scene text dataset, including horizontal, multi-oriented and curved scene texts. This dataset consists of 1,000 images for training and 500 images for testing. The annotation is mainly line-level and labeled by a polygon with 14 key points.

Total-Text [55] is also an arbitrary-shape scene text dataset, in which the training set has 1,255 images and the testing set

1520-9210 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: TU Delft Library. Downloaded on May 06,2021 at 19:28:26 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3073575, IEEE Transactions on Multimedia

TABLE I
ABLATION STUDIES OF SADA AND SSC ON THE DATASET *CTW1500*.

Method	IOU@0.5			IOU@0.7			Training time
	R (%)	P (%)	F (%)	R (%)	P (%)	F (%)	
MaskRCNN-alt [50]	**84.2**	80.9	82.5	69.6	71.0	70.3	$\sim$ 10h
Baseline	77.2	81.5	79.3	64.0	67.5	65.7	$>$ 30h
Ours (+SADA)	81.3	86.2	83.7	68.0	72.1	70.0	$\sim$ 4h
Ours (+SADA+SSC)	81.7	**87.2**	**84.4**	**70.0**	**74.7**	**72.3**	$\sim$ 4.5h

TABLE II
COMPARISONS BETWEEN SADA AND OTHER DATA AUGMENTATION STRATEGIES. THE EXPERIMENTS ARE CONDUCTED ON *CTW1500*.

Data	Method	IOU@0.5			Training Time
		R (%)	P (%)	F (%)	
Synthetic	SynthText [12]	**83.9**	81.4	82.6	$>$ 100h
Real	Bootstrapping [24]	82.5	80.6	81.5	$\sim$ 10h
	SNIPER [51]	80.0	84.8	82.3	$\sim$ 4h
	TDA-S	77.2	81.5	79.3	$>$ 30h
	TDA-R	75.9	80.4	78.1	$>$ 30h
	TDA-C	82.1	77.3	79.7	$\sim$ 3h
	Our SADA	81.3	**86.3**	**83.7**	$\sim$ 4h

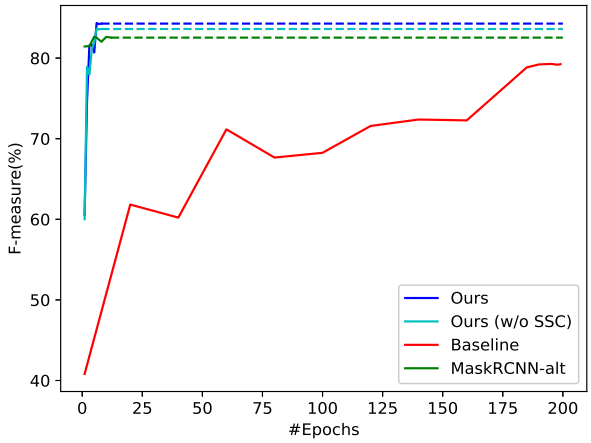


Fig. 5: Comparisons of the changes in *F-measure* against the training epoch on the dataset CTW1500. The dashed lines denote the expected changes. Best view in color.

has 300 images. All text instances are annotated by a world-level polygon with the unfixed number of key points.

ArT [56] is a larger arbitrary-shape scene text dataset. There are 10,166 images in total, including 5,603 training images and 4,563 testing images. The location of the scene text is annotated by the unfixed number of key points. This dataset not only contains English scene texts, but also involves lots of Chinese scene texts like that in RCTW-17 [57].

ICDAR2015 [58] is a multi-oriented scene text dataset. It contains 1,000 training images and 500 testing images. The annotation of each text in the image has 8 coordinates to enclose the text in a clockwise way, which makes the annotation word-level and polygonal.

ICDAR2013 [10] is a horizontal scene text dataset, which only contains horizontal or nearly horizontal text instances. In this dataset, there are 229 training images and 233 testing images. The word-level ground-truth bounding box is annotated by the top-left and bottom-right points.

COCO-Text [59] is the largest dataset for the localization task of the Latin scene text currently. The whole dataset contains 63,686 images with more than 17k text instances, in which $43,686$ images are selected as the training set, while $10,000$ images serve as the testing set and the rest is used for the validation. The annotation has two types, which are similar to that of *ICDAR2015* and *ICDAR2013*.

To evaluate the detection performances on CTW1500, ArT, ICDAR2015, and COCO-Text, we follow the IOU protocol provided in [11], [56], [58], [59], respectively. When evaluating on ICDAR2013, and Total-Text, the DetEval protocol provided in [10], [55] is employed.

B. Implementation Details

In experiments, we have fixed K, S, D_1, D_2, D_3, ρ_1, ρ_2, d and z to 512, 16, 1024, 2048, 256, 7, 14, 28 and 3, respectively. The base scales and aspect ratios of the anchors are set to (4, 6, 8, 10) and (0.5, 1, 2), respectively. In the step of training, we empirically set the resized scales $\{S_i \mid i$=1, 2, 3$\}$ to $\{$(512, 512), (800, 1280), (1400, 2000)$\}$ and the minimum edge ranges $\{R_i \mid i$=1, 2, 3$\}$ to $\{$(0, 100), (30, 160), (120, $+\infty$)$\}$. The parameters of the backbone network (ResNet-101) are initialized by the public model pre-trained on OpenImage [60]. The parameter initialization of the deformable convolution and DPS-ROIpooling is the same as that in [49]. The parameters in other layers are initialized as those in [50]. The whole model adopts an end-to-end training strategy, which is trained for 6 epochs. The learning rate is initially fixed to 0.015 and multiplied by 0.1 after 5 epochs. Other hyper-parameters in the framework are following those settings in [51]. In the testing, $\{S_i \mid i$=1, 2, 3$\}$ are set to $\{$(400, 600), (600, 800), (1000, 1400)$\}$ for CTW1500, Total-Text and ArT, and $\{$(400, 600), (800, 1200), (1400, 2000)$\}$ for ICDAR2013, ICDAR2015 and COCO-Text. The confidence threshold τ_c is set to 0.9 for CTW1500, Total-Text, ArT and ICDAR2013, and 0.65 for ICDAR2015 and COCO-Text, which are tuned by grid search on the training set as [8].

To compare with previous methods as fair as possible, we modify our model from four aspects. Firstly, we use ResNet-50 pre-trained on ImageNet [61] as the backbone instead of ResNet-101 pre-trained on OpenImage [60]. Secondly, we replace the deformable convolution with the standard convolution and the deformable position-sensitive ROI pooling with the ROIAlign [50]. Thirdly, we integrate the pyramid feature network (FPN) [62] into our model like that in Mask-RCNN [50], in that most existing methods also use FPN (or FPN variants). Finally, we only utilize a single scale in the inference. The testing scale is set to (600, 800) for CTW1500, Total-Text and ArT, to (1200, 2000) for ICDAR2015, and to (960, 1400) for ICDAR2013 and COCO-Text.

The proposed method is implemented based on the MXNet [63] framework. All experiments are carried out on a workstation with a 1.70 GHz Intel(R) Xeon(R) E5-2609 CPU, an NVIDIA GTX 1080Ti GPU, and 64G RAM.

C. Exploration of Proposed Modules

1) ***Baseline settings****:* The baseline is designed based on our proposed framework without the scale-aware data augmentation (SADA) and shape similarity constraint (SSC). The multi-scale training strategies and the training hyper-parameters follow the baseline settings in [33]. Specifically, the short

1520-9210 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: TU Delft Library. Downloaded on May 06,2021 at 19:28:26 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3073575, IEEE Transactions on Multimedia

TABLE III
INFLUENCES OF GAUSSIAN FILTER SIZE IN SSC. THE EXPERIMENTS ARE CONDUCTED ON THE DATASET *CTW1500*.

Method	*IOU*@0.7		
	R (%)	P (%)	F (%)
w/o SSC	68.0	72.1	70.0
SSC (z=3)	**70.0**	**74.7**	**72.3**
SSC (z=7)	68.2	74.9	71.4
SSC (z=11)	68.4	71.9	70.1
SSC (z=17)	68.9	71.7	70.3
SSC (z=23)	68.9	72.5	70.7
SSC (z=27)	68.6	73.2	70.8
SSIM Loss	69.0	73.0	71.0
Dice Loss	69.3	73.9	71.5

edges of the training images are randomly resized to three scales (640, 720, 800), and then each image is randomly flipped with a probability of 0.5. The whole model has been trained for 200 epochs in an end-to-end manner. The learning rate multiplies 0.1 after 180 epochs. We also train Mask-RCNN [50] that incorporates the feature pyramid network (FPN) [62] and utilizes the alternate training manner (denoted as 'MaskRCNN-alt'). Both RPN-1 and RPN-2 are trained for 6 epochs. Both MRCNN-1 and MRCNN-2 are trained for 12 epochs. The learning rate is multiplied by 0.1 after 5 and 11 epochs for RPN and MRCNN. The backbone also employs ResNet-101, which is initialized by the model pre-trained on the dataset OpenImage as well. The other hyper-parameter settings also follow the baseline settings in [33] for better detecting scene text. When testing, the scales of the input image are also the same as ours.

*2) **The influence of SADA**:* In Table I, IOU@0.5 and IOU@0.7 denote the evaluation protocols [11] where the IOU overlap threshold is set to 0.5 and 0.7. Experimental results have shown that, compared with the baseline, our SADA has a significant improvement of 4.7% and 4.4% in *Precision* and *F-measure*, under IOU@0.5. When using a stricter evaluation protocol IOU@0.7, the *Precision* and *F-measure* have still increased by 4.6% and 4.3%. When comparing with MaskRCNN-alt [50], SADA also achieves better *Precision* (+5.3%) and *F-measure* (+1.2%) under IOU@0.5. The *Recall* of MaskRCNN-alt is higher than our model with SADA, which mainly benefits from the influence of feature pyramid network (FPN) [62]. Besides, our method only costs about 4 hours to train the model in an end-to-end manner, which is obviously superior to the baseline. Fig. 5 also reveals that our model can achieve faster convergences and better performances. There are two reasons for this: i) SADA generates the image crops with lower resolutions to create the minibatch. These image crops can enlarge the minibatch size and discard some useless background regions in the original images. ii) SADA considers specific-scale scene texts in each image crop and makes full use of the valid text parts, which help to learn the diversities of training samples in each iteration.

To further verify the superiority of SADA, we compare it with the specially-designed data augmentation techniques achieved by the synthetic data and the real data. Some excellent researches [12], [45]–[47] devote to synthesizing scene text images. We just use the synthetic dataset SynthText [12] to pre-train our model one epoch as most existing methods, and then fine-tune on CTW1500. As shown in Table II, the

(a) (b)

Fig. 6: Examples of qualitative comparisions without SSC (a) and with SSC (b). Red denotes the detection result. Green is the ground-truth.

experimental results indicate that our SADA can achieve better performance, and especially save much training time. When comparing with [24] that augments training samples with real text parts, our SADA has improved the Precision of 5.7% and F-measure of 2.2%, respectively, as presented in Table II. When comparing with [51] that utilizes the real data to perform the efficient multi-scale training, the experimental results in Table II also show that our SADA works better in performance.

In addition, the traditional naive data augmentation techniques will also generate real data. Here, we mainly explore three kinds of traditional data augmentations through randomly scaling, rotating and cropping images (denoted as TDA-S, TDA-R and TDA-C), respectively, in that they can also affect the scales of scene texts. TDA-S is actually our baseline, which resizes the short edges of images to three scales (640, 720, 800). TDA-R randomly rotates the image from 0° to 360° with the interval of 10°. TDA-R only uses the single scale (600, 1000), which denotes the short edge is set to 600 while the long edge is not more than 1,000. Other settings of TDA-R follow those in TDA-S. TDA-C randomly crops the image with a single scale (600, 1000) into 512×512 patches like that in our SADA. The difference is that TDA-C allows all scene texts in each image crop to participate in the training. TDA-C trains 12 epochs and decays the learning rate by 0.1 at 10 epochs. All methods will randomly flip training images with a probability of 0.5. As shown in Table II, our SADA is significantly superior to TDA-S and TDA-R in both performance and training time. Although TDA-C can achieve faster convergence than our SADA, it is less effective than ours in performance (79.7% vs. 83.7%).

*3) **The influence of SSC**:* As shown in Table I, after SSC is incorporated with SADA to train the model, it can localize the scene text more accurately. Under the evaluation protocol IOU@0.7, it promotes 2.0%, 2.6% and 2.3% in terms of *Recall*, *Precision* and *F-measure*, respectively, compared with SADA. It also achieves the improvements of 1.0% and

1520-9210 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: TU Delft Library. Downloaded on May 06,2021 at 19:28:26 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3073575, IEEE Transactions on Multimedia

TABLE IV
F-MEASURE (%) UNDER DIFFERENT OVERLAP THRESHOLDS. THE EXPERIMENTS ARE CONDUCTED ON *CTW1500*. † DENOTES THE RESULTS FROM [34]. VALUES COLORED IN RED INDICATE THE DROPS VERSUS THE PERFORMANCE UNDER IOU@0.5.

Method	IOU			
	0.5	0.6	0.7	0.8
CTD-CLOC [11] †	73.4	64.3 (↓9.1)	46.6 (↓26.8)	19.5 (↓53.9)
Mask-TTD [34]	79.4	71.3 (↓8.1)	59.5 (↓19.9)	35.3 (↓44.1)
Ours	84.4	81.0 (↓3.4)	72.3 (↓12.1)	49.8 (↓34.6)

TABLE V
INFLUENCE OF BACKBONE. THE EXPERIMENTS ARE CONDUCTED ON THE DATASET *CTW1500*.

Backbone	IOU@0.5		
	R (%)	P (%)	F (%)
Res101 (OpenImage)	**81.7**	**87.2**	**84.4**
Res101 (ImageNet)	81.2	86.5	83.8
Res50 (ImageNet)	80.7	85.4	83.0
Res50 (ImageNet) w/o Deform	79.5	86.0	82.6
Res50-FPN (ImageNet) w/o Deform	80.4	86.2	83.2

0.7% in *Precision* and *F-measure* under IOU@0.5. When comparing with MaskRCNN-alt, our method has significantly improved the *Precision* of 6.3% and 3.7% under IOU@0.5 and IOU@0.7, respectively. In SSC, different Gaussian weighting filter size z captures various contexts for each spatial position, which will affect the boundary localization of scene texts. As shown in Table III, when $z = 3$, it achieves the best *F-measure* under IOU@0.7. The reason may be ascribed that the Gaussian filter is more capable of learning the boundary information when z is fixed to 3. Besides, SSC models the global shape structure of scene texts and backgrounds. It can effectively avoid detecting the text parts of the entire scene text and complex backgrounds, and relieve the false negatives. The qualitative detection results with and without SSC are presented in Fig. 6.

When our model employs the original SSIM loss used in BASNet [64] for high-quality salience object segmentation, the experimental results in Table III show that it will decrease by 1.0%, 1.7% and 1.3% in Recall, Precision and F-measure, respectively, compared with our SSC. It is because SSC can capture the shape structures of both scene texts and backgrounds under dual supervisions. Besides, when training in an end-to-end manner, the gradient of SSC will also affect the learning of localization of proposals. We further utilize the dice loss as [9] instead of our SSC to learn global-aware scene text boundaries in our model. Experimental results also demonstrate the effectiveness of our SSC, as shown in Table III. It is because our SSC is sensitive to the change of the scene text boundaries when a scene text is predicted as multiple text parts, but the dice loss is not aware of such changes when the text parts are very near.

Different from our SSC that purses more accurate localization of arbitrary-shape scene texts from the perspective of the loss function, Mask-TTD [34] exploits a tightness prior and the text frontier learning to enhance the pixel-wise mask prediction. However, as shown in Table IV, our SSC can achieve better performance under stricter overlap thresholds. When the overlap threshold ranges from 0.5 to 0.8, the drops of F-measure for our SSC are also lower.

4) ***The influence of backbone****:* To better analyze the performance of our proposed method, we further utilize different backbones to evaluate our model. As shown in Table V, when we utilize the ResNet-101 pre-trained on OpenImage [60] as the backbone, our method achieves the improvements of 0.5%, 0.7% and 0.6% in terms of *Recall*, *Precision* and *F-measure*, respectively, compared with that pre-trained on ImageNet [61]. When ResNet-50 pre-trained on ImageNet is employed as the backbone, it will decrease the *Recall* of 0.5%, *Precision* of 1.1% and *F-measure* of 0.8%, respectively, compared with ResNet-101 pre-trained on ImageNet. After we replace the deformable convolution with the standard convolution and the deformable position-sensitive ROI pooling with the ROIAlign [50], our model has achieved the Recall of 79.5%, Precision of 86.0% and F-measure of 82.6%, respectively. Further, we integrate the pyramid feature network (FPN) [62] with the backbone like that in Mask-RCNN [50] and only use a single scale in the inference. In Table V, the experimental results indicate that FPN can boost the F-measure of 0.6%, which is better than the multi-scale testing strategy for detecting different-scale scene texts.

TABLE VI
QUANTITATIVE COMPARISONS AMONG DIFFERENT NMS TYPES IN THE POST-PROCESSING. THE EXPERIMENTS ARE CONDUCTED ON *CTW1500*.

NMS Type	IOU@0.5		
	R (%)	P (%)	F (%)
BNMS	80.8	85.3	83.0
S-MNMS	**82.3**	85.1	83.7
MNMS	81.7	**87.2**	**84.4**

5) ***The influence of mask-level NMS****:* Since the text parts participate in the training, our model will also detect some text parts. To filter such redundant text parts, we utilize the mask-level NMS (MNMS), which is superior to the the box-level NMS (BNMS) [48] and standard mask-level NMS (S-MNMS) [50]. As shown in Table VI, the quantitative results have demonstarted that S-MNMS and MNMS can effectively improve the *Recall*, compared with BNMS. It is because BNMS will filter the highly-overlapped axis-aligned bounding boxes even though their corresponding masks are separated, which is illustrated in the first and second row of Fig. 7. Besides, as shown in Table VI, MNMS mainly increases the *Precison*, compared with BNMS and S-MNMS. It can be ascribed to the reason that MNMS is effective in removing the detected text parts in the whole text instance. Three samples in Fig. 7 also qualitatively illustrate the superiority of MNMS.

D. Generalization Ability

To verify the robustness in generalizing to unseen datasets, we adopt the cross-dataset evaluation rules as existing methods. That is, we evaluate our model by training on one dataset and testing on another dataset. We divide these evaluations into three situations.

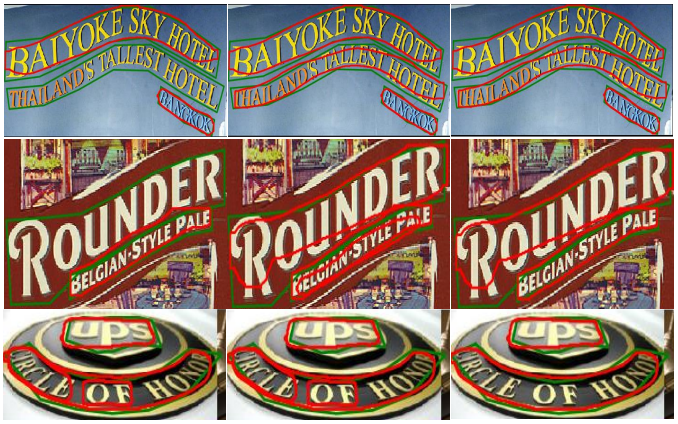
The first is evaluating the generalization abilities between arbitrary-shape text datasets (*CTW1500* and *Total-Text*). Specifically, we train our model on the training set of *CTW1500* and then test our model on the testing set of *Total-Text*, or our model is trained on the training set of *Total-Text* and then tested on the testing set of *CTW1500*. We adopt two evaluation protocols IOU@0.5 and DetEval as [35] for comprehensive evaluation. As shown in Table VII,

1520-9210 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: TU Delft Library. Downloaded on May 06,2021 at 19:28:26 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3073575, IEEE Transactions on Multimedia

TABLE VII
EXPLORATION OF GENERALIZATION ABILITY BETWEEN THE ARBITRARY-SHAPE SCENE DATASETS. † DENOTES THE RESULTS FROM [35]. ∗ MEANS MULTI-SCALE TESTING RESULTS.

Detector	Training set → Testing set											
	CTW1500 → *Total-Text*						*Total-Text* → *CTW1500*					
	IoU@0.5			*DetEval*			*IoU*@0.5			*DetEval*		
	R(%)	P(%)	F(%)	R(%)	P(%)	F(%)	R(%)	P(%)	F(%)	R(%)	P(%)	F(%)
CTD-TLOC [11]†	25.9	42.5	32.2	34.5	44.1	38.7	50.5	37.4	43.0	42.9	44.1	43.5
TFE-PRPA-BCTS [35]	31.0	55.5	39.8	47.2	66.6	55.2	**57.8**	35.1	43.6	65.5	62.1	63.7
Ours (Res50-FPN-ImageNet)	32.6	57.1	41.5	66.7	71.2	68.9	56.2	36.7	44.4	67.2	68.1	67.6
Ours (Res101-OpenImage)*	**34.2**	**59.0**	**43.3**	**68.9**	**74.8**	**71.7**	57.5	**37.9**	**45.7**	**70.5**	**72.1**	**71.3**



(a) (b) (c)

Fig. 7: Qualitative detection results from different NMS types. (a) Box-level NMS. (b) Standard Mask-level NMS. (c) Mask-level NMS used in our method. Green denotes the ground-truth. Red is the detection result.

our method achieves excellent generalization performances, which significantly outperform the generalization abilities of the pioneering arbitrary-shape detector CTD-TLOC [11].

The second is evaluating the generalization abilities between the arbitrary-shape text dataset (*Total-Text*) and the multi-oriented text dataset (*ICDAR2015*). Specifically, we train our model on the training set of *Total-Text* and then test our model on the testing set of *ICDAR2015*. Meanwhile, our model is also trained on the training set of *ICDAR2015* and then tested on the testing set of *Total-Text*. As shown in Table VIII, the generalization ability of our model is superior to all other well-known detectors.

The third is that our model is first trained on the training set of the multi-oriented text dataset *ICDAR2015*, and then is tested on the testing set of the horizontally-oriented text dataset *COCO-Text*[1]. As illustrated in Table IX, compared with most excellent detectors, our model has achieved better generalization performance. Although Mask-TextSpotter [36] adopts the synthetic data [12] to pre-train the model, utilize more real-world data to fine-tune the model and integrate the recognition network to promote the detection performance, the *F-measure* of our model just has decreased by 0.1% in a single-scale testing.

[1]*COCO-Text* has two annotation types, but it utilizes the annotations like ICDAR2013 when evaluating the performance of the model. Therefore, we take this dataset as a horizontally-oriented scene text dataset here.

TABLE VIII
EXPLORATION OF GENERALIZATION ABILITY BETWEEN THE ARBITRARY-SHAPE SCENE TEXT DATASET *Total-Text* AND THE MULTI-ORIENTED SCENE TEXT DATASET *ICDAR2015*. † DENOTES THE RESULTS FROM [8]. ∗ MEANS MULTI-SCALE TESTING RESULTS.

Detector	Training set → Testing set					
	ICDAR2015 →*Total-Text*			*Total-Text* → *ICDAR2015*		
	DetEval			*IoU*@0.5		
	R(%)	P(%)	F(%)	R(%)	P(%)	F(%)
SegLink [20] †	33.2	35.6	34.0	–	–	–
EAST [7] †	43.1	49.0	45.9	–	–	–
PixelLink [39] †	52.7	53.5	53.1	–	–	–
TextSnake [8]	67.9	61.5	64.6	–	–	–
TextField [28]	65.2	61.5	63.3	66.0	77.1	71.1
PAN [41]	57.8	72.0	64.1	65.5	**77.6**	71.1
Ours (Res50-FPN-ImageNet)	56.5	79.2	66.0	70.3	75.6	72.9
Ours (Rest101-OpenImage)*	**58.2**	**81.5**	**67.9**	**72.7**	77.5	**75.0**

TABLE IX
EXPLORATION OF GENERALIZATION ABILITY FROM THE MULTI-ORIENTED SCENE TEXT DATASET *ICDAR2015* TO THE HORIZONTALLY-ORIENTED SCENE TEXT DATASET *COCO-Text*. ∗ INDICATES MULTI-SCALE TESTING RESULTS. △ DENOTES TRAINING ON *COCO-Text*.

Detector	*ICDAR2015* → *COCO-Text*		
	IoU@0.5		
	R(%)	P(%)	F(%)
EAST [7]	32.4	50.4	39.5
SSTD [22]	31.0	46.0	37.0
WordSup [25]*	30.9	45.2	36.8
RRD [16]	57.0	64.0	61.0
TextBoxes++ [18]△*	56.7	60.9	58.7
Mask-TextSpotter [36]	58.3	66.8	62.3
PS-COI [32]△	39.0	61.0	47.0
Ours (Res50-FPN-ImageNet)	57.2	68.1	62.2
Ours (Res101-OpenImage)*	**59.1**	**70.3**	**64.2**

E. Comparisons with Related Methods

Comparisons On CTW1500: As shown in Table X, when not using the synthetic text dataset *SynthText* [12] to pre-train the model, our model with single-scale testing is superior to all state-of-the-art methods. Compared with existing methods that employ *SynthText*, our model is still better than most of them. We further utilize the ResNet-101 pre-trained on OpenImage as the backbone and multi-scale testing strategy, our method can achieve better performance. The qualitative detection results are presented in the first row of Fig. 8.

Comparisons On Total-Text: As shown in Table XI, the Recall of our proposed method outperforms the state-of-the-art methods that do not pre-train on the synthetic dataset *SynthText* [12]. Specifically, our model improves the *Recall* of 1.8% compared with the detector PAN [41] in a single-scale testing. When we utilize ResNet-101 pre-trained on OpenImage to initialize our model and test the model using multiple scales, our method also works better than other methods that use multi-scale testing strategies and pre-train on *SynthText*. The second row of Fig. 8 shows the qualitative detection results.

Comparisons On ArT: For this dataset, we submit our detection results to the website and evaluate our proposed

1520-9210 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: TU Delft Library. Downloaded on May 06,2021 at 19:28:26 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3073575, IEEE Transactions on Multimedia

TABLE X
COMPARISONS WITH RELATED WORKS ON *CTW1500*. 'Using SynthText' MEANS USING THE SYNTHETIC DATASET TO PRE-TRAIN THE MODEL. THE UNDERLINED AND THE BOLD DENOTE THE OPTIMAL VALUES AMONG THE METHODS PRE-TRAINED WITH AND WITHOUT USING 'SynthText'. * INDICATES MULTI-SCALE TESTING RESULTS.

Method	Publication	Using SynthText	R (%)	P (%)	F (%)
TextSnake [8]	ECCV'18	√	*85.3*	67.9	75.6
PSENet [9]	CVPR'19	√	79.7	84.8	82.2
MSR [29]*	IJCAI'19	√	78.3	85.0	81.5
CRAFT [27]	CVPR'19	√	81.1	86.0	83.5
PAN [41]	ICCV'19	√	81.2	*86.4*	*83.7*
LOMO [37]*	CVPR'19	√	76.5	85.7	80.8
SAE [40]	CVPR'19	√	77.8	82.7	80.1
SAST [30]*	MM'19	√	81.7	81.2	81.5
ICG [21]	PR'19	√	79.8	82.8	81.3
TextField [28]	TIP'19	√	79.8	83.0	81.4
ATRR [26]	CVPR'19	×	80.2	80.1	80.1
PAN [41]	ICCV'19	×	77.7	84.6	81.0
CTD-CLOC [11]	PR'19	×	69.8	77.4	73.4
AB-LSTM [31]	TOMM'19	×	81.6	83.0	82.3
Mask-TTD [34]	TIP'20	×	79.0	79.7	79.4
Ours (Res50-FPN-ImageNet)	—	×	80.4	86.2	83.2
Ours (Res101-OpenImage)*	—	×	**81.7**	**87.2**	**84.4**

TABLE XI
COMPARISONS WITH RELATED WORKS ON *Total-Text*. 'Using SynthText' MEANS USING THE SYNTHETIC DATASET TO PRE-TRAIN THE MODEL. THE UNDERLINED AND THE BOLD DENOTE THE OPTIMAL VALUES AMONG THE METHODS PRE-TRAINED WITH AND WITHOUT USING 'SynthText'. * INDICATES MULTI-SCALE TESTING RESULTS.

Method	Publication	Using SynthText	R (%)	P (%)	F (%)
TextSnake [8]	ECCV'18	√	74.5	82.7	78.4
Mask-TextSpotter [36]	TPAMI'19	√	75.4	81.8	78.5
TextField [28]	TIP'19	√	79.9	81.2	80.6
ICG [21]	PR'19	√	80.9	82.1	81.5
MSR [29]*	IJCAI'19	√	74.8	83.8	79.0
SPCNet [33]	AAAI'19	√	*82.8*	83.0	82.9
LOMO [37]*	CVPR'19	√	79.3	87.6	83.3
PSENet [9]	CVPR'19	√	78.0	84.0	80.9
SAST [30]*	MM'19	√	75.5	85.6	80.2
PAN [41]	ICCV'19	√	81.0	*89.3*	*85.0*
TFE-PRPA-BCTS [35]	TMM'20	√	78.6	84.6	81.5
DeconvNet-Text [55]	ICDAR'17	×	33.0	44.0	36.0
ATRR [26]	CVPR'19	×	76.2	80.9	78.5
PAN [41]	ICCV'19	×	79.4	**88.0**	83.5
CTD-CLOC [11]	PR'19	×	71.0	74.0	73.0
AB-LSTM [31]	TOMM'19	×	78.2	78.9	78.5
TFE-PRPA-BCTS [35]	TMM'20	×	74.7	83.5	78.9
Mask-TTD [34]	TIP'20	×	74.5	79.1	76.7
Ours (Res50-FPN-ImageNet)	—	×	81.2	85.4	83.2
Ours (Res101-OpenImage)*	—	×	**82.6**	86.7	**84.6**

model online. As shown in Table XII, when using the ResNet-50 pre-trained on ImageNet as our backbone and integrating the feature pyramid network (FPN) [62], our method achieves the *Recall* of 68.0%, *Precision* of 81.3% and *F-measure* of 74.1% in a single-scale testing, which outperforms all other arbitrary-shape detectors listed in Table XII. When we employ a stronger backbone (ResNet-101 pre-trained on OpenImage) and use the multi-scale testing strategy, our model achieves the best *Precision* and *F-measure*.

Comparisons On ICDAR2015: In Table XIII, without utilizing *SynthText* [12] to pre-train the model, our method can achieve the best performance when adopting a single-scale or multi-scale testing strategy. Besides, our model using ResNet-101 pre-trained on OpenImage as backbone also outperforms most well-known multi-oriented detectors (e.g.,EAST [7], RRD [16] and TextBoxes++ [18]) in a multi-scale testing strategy. The third row of Fig. 8 shows the qualitative detection results of our method.

Comparisons On ICDAR2013: Similar to PixelLink [39], we utilize the model trained on *ICDAR2015* to initialize our model, and then fine-tune on the training set of *ICDAR2013*.

TABLE XII
COMPARISONS WITH RELATED METHODS ON THE DATASET *ArT*. † MEANS THE EVALUATION RESULTS FROM THE COMPETITION LEADERBOARD ABOUT ARBITRARY-SHAPE SCENE TEXT DETECTION [56].

Detector	IOU@0.5		
	R (%)	P (%)	F (%)
MSR [29] †	0.46	0.55	0.50
TextCohesion †	43.7	68.1	53.2
TMIS †	53.5	**86.2**	66.0
MFTD †	63.1	72.1	67.3
CCISTD †	60.7	81.2	69.5
QAQ †	63.5	83.8	72.2
CRAFT [27] †	68.9	77.3	72.9
CLTDR †	65.9	82.6	73.3
Ours (Res50-FPN-ImageNet)	68.0	81.3	74.1
Ours (Res101-OpenImage)*	**69.5**	83.3	**75.8**

TABLE XIII
COMPARISIONS WITH RELATED WORKS ON *ICDAR2015*. 'Using SynthText' MEANS USING THE SYNTHETIC DATASET TO PRE-TRAIN THE MODEL. THE UNDERLINED AND THE BOLD DENOTE THE OPTIMAL VALUES AMONG THE METHODS PRE-TRAINED WITH AND WITHOUT USING 'SynthText'. * INDICATES MULTI-SCALE TESTING RESULTS.

Method	Publication	Using SynthText	R (%)	P (%)	F (%)
SegLink [20]	CVPR'17	√	76.8	73.1	75.0
WordSup [25]*	ICCV'17	√	77.0	79.3	78.2
TextCorner [17]*	CVPR'18	√	79.7	*89.5*	84.3
RRD [16]*	CVPR'18	√	80.0	88.0	83.8
TextSnake [8]	ECCV'18	√	80.4	84.9	82.6
TextBoxes++ [18]*	TIP'18	√	78.5	87.8	82.9
SPCNet [33]	AAAI'19	√	85.8	88.7	87.2
PSENet [9]	CVPR'19	√	85.2	89.3	87.2
PAN [41]	ICCV'19	√	81.9	84.0	82.9
SAST [30]*	MM'19	√	*87.3*	87.6	*87.4*
SAE [40]	CVPR'19	√	85.0	88.3	86.6
MSR [29]*	IJCAI'19	√	78.4	86.6	82.3
TextField [28]	TIP'19	√	80.5	84.3	82.4
EAST [7]*	CVPR'17	×	78.3	83.3	80.7
DeepReg [23]	ICCV'17	×	80.0	82.0	81.0
PixelLink [39]	AAAI'18	×	82.7	82.9	82.3
ITN [19]	CVPR'18	×	74.1	85.7	79.5
VIS [45]	ECCV'18	×	77.2	87.1	81.9
RRPN [5]	TMM'18	×	73.0	82.0	77.0
GA-DAN [47]	ICCV'19	×	81.6	85.6	83.5
PAN [41]	ICCV'19	×	81.9	84.0	82.9
Ours (Res50-FPN-ImageNet)	—	×	81.3	87.2	84.1
Ours (Res101-OpenImage)*	—	×	**82.6**	**88.8**	**85.6**

As shown in Table XIV, when using the ResNet-101 pre-trained on OpenImage and the multi-scale testing strategy, our method achieves the best-second performance among the methods that do not utilize the *SynthText* [12] to pre-train the model. Although the performance of FEN [13] is superior to ours, it utilizes the self-collected data to train the model. In addition, FEN [13] is designed to detect the horizontal or nearly-horizontal scene text, but our method can detect the arbitrary-shape scene text. In Fig. 8, the fourth row displays the qualitative detection results of our model.

F. Runtime Analyses

In the inference stage, the runtime of our method is mainly influenced by the network inference and post processing. For different datasets *CTW1500*, *Total-Text*, *ICDAR2015* and *ICDAR2013*, the scale of the input image and the number of text instances will lead to different runtime, as shown in Table XV. We calculate the average time per image in each dataset based on a workstation with one NVIDIA GTX 1080Ti GPU and the Intel i7 CPU. Besides, we do not utilize any parallel procedures to accelerate the post processing. The reported runtime in Table XV has shown that our model can detect the arbitrary-shape scene texts in decent runtime.

1520-9210 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: TU Delft Library. Downloaded on May 06,2021 at 19:28:26 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3073575, IEEE Transactions on Multimedia

Fig. 8: Qualitative detection results of our method. The row from top to bottom denotes the detection results on *CTW1500*, *Total-Text*, *ICDAR2015* and *ICDAR2013*, respectively. Red is the detection result. Green represents the ground-truth.

TABLE XIV
COMPARISONS WITH RELATED WORKS ON *ICDAR2013*. 'Using SynthText' MEANS USING THE SYNTHETIC DATASET TO PRE-TRAIN THE MODEL. THE UNDERLINED AND THE BOLD DENOTE THE OPTIMAL VALUES AMONG THE METHODS PRE-TRAINED WITH AND WITHOUT USING 'SynthText'. * INDICATES MULTI-SCALE TESTING RESULTS.

Method	Publication	Using SynthText	R (%)	P (%)	F (%)
TextBoxes [4]*	AAAI'17	√	83.0	89.0	86.0
SegLink [20]	CVPR'17	√	83.0	87.7	85.3
TextCorner [17]*	CVPR'18	√	84.4	92.0	88.0
TextBoxes++ [18]*	TIP'18	√	86.0	92.0	89.0
SPCNet [33]	AAAI'19	√	*90.5*	*93.8*	*92.1*
CTPN [6]	ECCV'16	×	83.0	93.0	87.7
DeepReg [23]	ICCV'17	×	81.0	92.0	86.0
SSTD [22]	ICCV'17	×	86.0	89.0	88.0
TSM [14]	TMM'17	×	67.0	81.0	73.0
FEN [13]*	AAAI'18	×	**90.0**	**94.7**	**92.3**
PixelLink [39]*	AAAI'18	×	87.5	88.6	88.1
SSFT-DLRC [15]	TMM'18	×	86.1	91.1	88.5
Ours (Res50-FPN-ImageNet)	—	×	85.1	89.8	87.4
Ours (Res101-OpenImage)*	—	×	86.2	91.1	88.6

TABLE XV
RUNTIME OF THE PROPOSED METHOD. THE RUNTIME IS ACQUIRED WITH A SINGLE NVIDIA GTX 1080Ti GPU.

Dataset	Network inference	Post processing
CTW1500	1.61 s	0.09 s
Total-Text	1.31 s	0.16 s
ICDAR2015	1.36 s	0.12 s
ICDAR2013	1.46 s	0.17 s

G. Limitations

Our proposed method is capable of working well in many challenging scenarios, but there are still some failure cases. Firstly, although the mask-level NMS is successful in removing the detected text parts in the text instance, it fails to remain the valid text instances when the overlaps between text instances are large. As shown in Fig. 9 (a), the text instances (cyan) are filtered. Secondly, some extremely low-contrast scene texts are not detected, as shown in Fig. 9 (b). It is because the low-contrast scene text instances in the training set are rare, which is not helpful to the learning of deep neural networks. Although our data augmentation technique can increase the training samples, it focuses on the scales and hardly increases the low-contrast training samples. Besides, the examples in Fig. 9 (c) have illustrated that some very text-like backgrounds will not be filtered by our model. It may be ascribed that our model cannot learn the text-like patterns well. This flaw may be relieved by the hard example mining or the recognition network.

(a) (b) (c)

Fig. 9: Failure samples. Top samples are from the dataset *CTW1500*. Bottom samples come from the dataset *Total-Text*. Red denotes the detection result. Green is the ground-truth.

1520-9210 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: TU Delft Library. Downloaded on May 06,2021 at 19:28:26 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3073575, IEEE Transactions on Multimedia

V. Conclusion

In this paper, we have presented an end-to-end trainable framework to detect the arbitrary-shape scene text. In the framework, a novel scale-aware data augmentation technique is proposed to increase the diversity of training data, for faster convergence and better performance. Meanwhile, a novel shape similarity constraint is introduced to generate a more accurate localization for the arbitrary-shape scene text. These two proposed techniques can be seamlessly integrated into the training. Extensive experiments conducted on several public benchmarks have demonstrated the effectiveness, superiority and generalization ability of our proposed method. In the future, we are interested in improving the inference time of our proposed model and incorporating the scene text recognition network with our detection architecture for the end-to-end arbitrary-shape scene text spotting.

References

[1] A. F. Biten, R. Tito, A. Mafla, L. G. i Bigorda, M. Rusiñol, C. V. Jawahar, E. Valveny, and D. Karatzas, "Scene text visual question answering," in *ICCV*, 2019, pp. 4290–4300.

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

[3] Z. Hong, Y. Petillot, D. Lane, Y. Miao, and S. Wang, "TextPlace: Visual place recognition and topological localization through reading scene texts," in *ICCV*, 2019, pp. 2861–2870.

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

[5] 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.

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

[7] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: An efficient and accurate scene text detector," in *CVPR*, 2017, pp. 2642–2651.

[8] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "TextSnake: A flexible representation for detecting text of arbitrary shapes," in *ECCV*, 2018, pp. 19–35.

[9] 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 *CVPR*, 2019, pp. 9336–9345.

[10] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, and et al., "Icdar 2013 competition on robust reading," in *ICDAR*, 2013, pp. 1484–1493.

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

[12] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *CVPR*, 2016, pp. 2315–2324.

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

[14] X. Ren, Y. Zhou, J. He, K. Chen, X. Yang, and J. Sun, "A convolutional neural network-based chinese text detection algorithm via text structure modeling," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 506–518, 2017.

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

[16] M. Liao, Z. Zhu, B. Shi, G.-s. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *CVPR*, 2018, pp. 5909–5918.

[17] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in *CVPR*, 2018, pp. 7553–7563.

[18] 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.

[19] F. Wang, L. Zhao, X. Li, X. Wang, and D. Tao, "Geometry-aware scene text detection with instance transformation network," in *CVPR*, 2018, pp. 1381–1389.

[20] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *CVPR*, 2017, pp. 3482–3490.

[21] J. Tang, Z. Yang, Y. Wang, Q. Zheng, Y. Xu, and X. Bai, "SegLink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping," *Pattern Recognit.*, vol. 96, 2019.

[22] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in *ICCV*, 2017, pp. 3066–3074.

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

[24] C. Xue, S. Lu, and F. Zhan, "Accurate scene text detection through border semantics awareness and bootstrapping," in *ECCV*, 2018, pp. 370–387.

[25] H. Hu, C. Zhang, Y. Luo, Y. Wang, J. Han, and E. Ding, "WordSup: Exploiting word annotations for character based text detection," in *ICCV*, 2017, pp. 4950–4959.

[26] X. Wang, Y. Jiang, Z. Luo, C. Liu, H. Choi, and S. Kim, "Arbitrary shape scene text detection with adaptive text region representation," in *CVPR*, 2019, pp. 6449–6458.

[27] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *CVPR*, 2019, pp. 9365–9374.

[28] 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.

[29] C. Xue, S. Lu, and W. Zhang, "MSR: multi-scale shape regression for scene text detection," in *IJCAI*, 2019, pp. 989–995.

[30] P. Wang, C. Zhang, F. Qi, Z. Huang, M. En, J. Han, J. Liu, E. Ding, and G. Shi, "A single-shot arbitrarily-shaped text detector based on context attended multi-task learning," in *ACM-MM*, 2019, pp. 1277–1285.

[31] Z. Liu, W. Zhou, and H. Li, "AB-LSTM: Atention-based bidirectional lstm model for scene text detection," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 4, pp. 1–23, 2019.

[32] P. Cheng, Y. Cai, and W. Wang, "A direct regression scene text detector with position-sensitive segmentation," *IEEE Trans. Circuits Sys. Video Technology.*, vol. In Press, 2020.

[33] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao, and G. Li, "Scene text detection with supervised pyramid context network," in *AAAI*, 2019, pp. 9038–9045.

[34] Y. Liu, L. Jin, and C. Fang, "Arbitrarily shaped scene text detection with a mask tightness text detector," *IEEE Trans. Image Process.*, vol. 29, pp. 2918–2930, 2020.

[35] P. Dai, H. Zhang, and X. Cao, "Deep multi-scale context aware feature aggregation for curved scene text detection," *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 1969–1984, 2020.

[36] 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. In Press, 2019.

[37] 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 *CVPR*, 2019, pp. 10 552–10 561.

[38] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 161–184, 2021.

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

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

[41] 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 *ICCV*, 2019, pp. 8439–8448.

[42] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *AAAI*, 2020, pp. 13 001–13 008.

[43] H.-S. Fang, J. Sun, R. Wang, M. Gou, Y.-L. Li, and C. Lu, "Instaboost: Boosting instance segmentation via probability map guided copy-pasting," in *ICCV*, 2019, pp. 682–691.

[44] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," in *Workshop on Deep Learning, NeurIPS*, 2014.

[45] F. Zhan, S. Lu, and C. Xue, "Verisimilar image synthesis for accurate detection and recognition of texts in scenes," in *ECCV*, 2018, pp. 257–273.

[46] F. Zhan, H. Zhu, and S. Lu, "Spatial fusion GAN for image synthesis," in *CVPR*, 2019, pp. 3653–3662.

1520-9210 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: TU Delft Library. Downloaded on May 06,2021 at 19:28:26 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3073575, IEEE Transactions on Multimedia

[47] F. Zhan, C. Xue, and S. Lu, "GA-DAN: geometry-aware domain adaptation network for scene text detection and recognition," in *ICCV*, 2019, pp. 9104–9114.
[48] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015, pp. 91–99.
[49] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *ICCV*, 2017, pp. 764–773.
[50] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *ICCV*, 2017, pp. 2980–2988.
[51] B. Singh, M. Najibi, and L. S. Davis, "SNIPER: efficient multi-scale training," in *NeurIPS*, 2018, pp. 9333–9343.
[52] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
[53] Y. Dai, Z. Huang, Y. Gao, Y. Xu, K. Chen, J. Guo, and W. Qiu, "Fused text segmentation networks for multi-oriented scene text detection," in *ICPR*, 2018, pp. 3604–3609.
[54] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," in *SIGGRAPH*, 1987, pp. 163–169.
[55] C. K. Ch'ng and C. S. Chan, "Total-Text: A comprehensive dataset for scene text detection and recognition," in *ICDAR*, 2017, pp. 935–942.
[56] C. K. Chng, E. Ding, J. Liu, D. Karatzas, and et al., "ICDAR2019 robust reading challenge on arbitrary-shaped text - rrc-art," in *ICDAR*, 2019, pp. 1571–1576.
[57] B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. J. Belongie, S. Lu, and X. Bai, "ICDAR2017 competition on reading Chinese text in the wild (RCTW-17)," in *ICDAR*, 2017, pp. 1429–1434.
[58] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, and et al., "Icdar 2015 competition on robust reading," in *ICDAR*, 2015, pp. 1156–1160.
[59] R. Gomez, B. Shi, L. Gomez-Bigorda, and et al., "ICDAR2017 robust reading challenge on coco-text," in *ICDAR*, 2017, pp. 1435–1443.
[60] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, and et al, "OpenImages: A public dataset for large-scale multi-label and multi-class image classification," Dataset available from https://storage.googleapis.com/openimages/web/index.html, 2017.
[61] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.
[62] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017, pp. 936–944.
[63] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, "MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems," *arXiv preprint arXiv:1512.01274*, 2015.
[64] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jägersand, "BASNet: Boundary-aware salient object detection," in *CVPR*, 2019, pp. 7479–7489.

Pengwen Dai is currently a Ph.D. candidate with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. His current research interests include scene text detection and recognition.

Yang Li is a PostDoc at the Algorithmics Group of the Faculty of Engineering, Mathematics and Computer Science (EEMCS/EWI), Delft University of Technology. She received her B.E. and Ph.D. degrees in 2014 and 2019, from the Chongqing University and Tsinghua University of China, respectively. Her research interests include intelligent vehicles, reinforcement learning and its application in autonomous driving.

Hua Zhang is an associate professor with the Institute of Information Engineering, Chinese Academy of Sciences. He received the Ph.D. degrees in computer science from the School of Computer Science and Technology, Tianjin University, Tianjin, China in 2015. His research interests include computer vision, multimedia, and machine learning.

Jingzhi Li is currently pursuing the Ph.D. degree with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China. Her current research interests include face privacy, unbiased facial recognition, and machine learning.

Xiaochun Cao received the B.E. and M.E. degrees in computer science from Beihang University, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, USA. After graduation, he spent about three years at ObjectVideo Inc. as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. He has been a professor with the Institute of Information Engineering, Chinese Academy of Sciences, since 2012. He is also with the Peng Cheng Laboratory, Cyberspace Security Research Center, China, and the School of Cyber Security, University of Chinese Academy of Sciences, China. He is on the Editorial Boards of the IEEE Transactions on Image Processing, IEEE Transactions on Multimedia and IEEE Transactions on Circuits and Systems for Video Technology. In 2004 and 2010, he was the recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition.

1520-9210 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: TU Delft Library. Downloaded on May 06,2021 at 19:28:26 UTC from IEEE Xplore. Restrictions apply.