electronics

MDPI

Article

Exploring Style-Robust Scene Text Detection via Style-Aware Learning

Yuanqiang Cai [1,2], Fenfen Zhou [3,4,*] and Ronghui Yin [5]

[1] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China; caiyuanqiang@bupt.edu.cn
[2] School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
[3] Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China
[4] College of Robotics, Beijing Union University, Beijing 100027, China
[5] Guangdong Science & Technology Infrastructure Center, Guangzhou 510033, China; ronghuiyin1@gmail.com
* Correspondence: jqrfenfen@buu.edu.cn

Abstract: Although current scene text detectors achieve remarkable accuracy across different and diverse styles of datasets by fine-tuning models multiple times, these approaches are time-consuming and hinder model generalization. As such, exploring a training strategy that only requires training once on all datasets is a promising solution. However, the text-style mismatch poses challenges to accuracy in such an approach. To mitigate these issues, we propose a style-aware learning network (SLNText) for style-robust text detection in the wild. This includes a style-aware head to distinguish the text styles of images and a dynamic selection head to realize the detection of images with different text styles. SLNText is only trained once, achieving superior performance by automatically learning from multiple text styles and overcoming the style mismatch issue inherent in one-size-fits-all approaches. By using only one set of network parameters, our method significantly reduces the training consumption while maintaining a satisfactory performance on several styles of datasets. Our extensive experiments demonstrate that SLNText achieves satisfactory performance in several styles of datasets, showcasing its effectiveness and efficiency as a promising solution to style-robust scene text detection.

Keywords: scene text detection; style-aware learning; dynamic selection head; dynamic training strategy

Citation: Cai, Y.; Zhou, F.; Yin, R. Exploring Style-Robust Scene Text Detection via Style-Aware Learning. *Electronics* **2024**, *13*, 243. https://doi.org/10.3390/electronics13020243

Academic Editor: George A. Tsihrintzis

Received: 6 December 2023
Revised: 25 December 2023
Accepted: 28 December 2023
Published: 5 January 2024

1. Introduction

Scene text detection has become more and more popular with the development of computer vision. It has a wide range of practical applications across various styles and constitutes a crucial predecessor for, e.g., image and video retrieval, autonomous driving, scene understanding, retail analytics, and industrial automation. Scene text refers to the written content that naturally occurs within various real-world contexts. The fonts and styles used in scene text are influenced by the specific surroundings, leading to a diverse range of text appearances. The different scenes we encounter contribute to the unique styles of text. However, this variety in text styles presents a challenge when training detectors and expecting them to work well across different scenarios.

To improve the performance of scene text detectors in different scenarios, researchers put forward various scene text detection datasets [1–7] with specific characteristics. The existing algorithms [8–13] have good performances on different datasets, respectively, by fine-tuning and retraining on each benchmark with individual weight. In contrast to existing datasets predominantly characterized by a singular text style, real-world scenarios exhibit a variety of distinct text styles. Furthermore, the textual styles within the input image of application scenarios during the testing phase remain uncertain. Consequently,

the style-robust efficacy of extant scene text detectors within these intricate application scenarios faces more challenges.

To verify our claim, we take ICDAR 2013 [1], ICDAR 2015 [2], MSRA-TD500 [3], and CTW1500 [4] datasets as examples, and use the union of the four datasets as an application scenario containing various styles of text. Faced with this situation, the existing methods adopt two prevailing training strategies. The first strategy is to pre-train on a combined dataset followed by individual fine-tuning on each distinct dataset. This causes the algorithm to generate a set of parameters for each dataset, which makes the algorithm perform well on the respective dataset. However, this strategy can lead to the over-fitting of the corresponding dataset, so that the set of parameters of one dataset does not perform well on other datasets. Furthermore, during testing, the selection of appropriate parameter sets becomes infeasible due to the indeterminate text style within the combined dataset's input images.

Another strategy is to directly use the combined dataset for training, which can take into account the detection performance of different styles of text. However, distinct text styles have a large difference and will interfere with each other during the training stage, resulting in the detection performance of the detector on a certain dataset that is not as good as the performance of the method only training on the single dataset.

In this paper, we propose to devise a novel detection algorithm alongside a congruous training strategy capable of precision text detection across scenes characterized by diverse text styles. This algorithm needs to address two challenges. Style diversity in feature learning: The proposed model needs to acquire distinctive text style features while mitigating the inadvertent cross-contamination of feature parameters. Style discrimination and accurate prediction: During inference, how does the model distinguish disparate text styles and deliver precise text regions?

Inspired by the dynamic network [14,15], we designed a style-aware learning network for scene text detection named SLNText. It includes a style-aware head (SAH) and a dynamic selection head (DSH), where each pair of heads represents different text features based on the corresponding style. In contrast to previous methods with multi-time fine-tuning, the proposed SLNText only needs to train once for the shared weight with a dynamic head. With the assistance of a style-aware head, DSH can dynamically adjust the network parameters according to the visual features (e.g. scene type, text appearance) of the input image to alleviate the negative mutual influence from different styles, resulting in better detection performance. To efficiently train the proposed network, we propose a dynamic training strategy (DTS) that can train the proposed network with a small additional training cost compared to the training on the combined dataset. Extensive experiments on four datasets (i.e., ICDAR 2013, ICDAR 2015, MSRA-TD500, and CTW1500) support our analysis and conclusions. Compared with the baseline of the proposed method, the proposed algorithm and training strategy improve by at least 3.84% F-measure performance on the combined dataset.

The main contributions of this work can be summarized as:

- We propose a style-aware learning network for style-robust scene text detection named SLNText. By equipping the style-aware head (SAH) and the dynamic selection head (DSH), this reduces the style gap by dynamically adjusting the network parameters.
- A dynamic training strategy is proposed to train the proposed SLNText with little additional training cost.
- The proposed method achieves an improvement with a mean F-measure of 3.85% with the dynamic prediction heads, where each head represents different features from different styles.

2. Related Work

Due to the various fonts and contours in the text objects, coupled with the rich variety of the text scenes, the text style has become diverse, which has brought great challenges to detection methods. In this section, we conduct a comprehensive review of

existing approaches that tackle these challenges from two key perspectives: feature learning, enabling effective text discrimination in complex scenarios; and contour representation, facilitating precise bounding box delineation.

2.1. Feature Learning for Scene Text Detector

Early text detection algorithms based on deep learning mostly used VGG backbone networks [16] for multi-resolution feature extraction [17–19]. TextBox [17] utilizes the multi-scale characteristics of the backbone network itself to learn the feature information for the corresponding scale texts at different feature layers. With the introduction of the ResNet backbone network [20] and the feature pyramid network [21], researchers have habitually combined the two for the multi-scale feature extraction of text. Recent methods [22] integrate text knowledge into ResNet and feature pyramid networks through image-level text recognition tasks and then migrate the new backbone network to detection algorithms to improve the detection accuracy. Reference [23] used ResNet and a U-shaped network similar to a feature pyramid to construct a multi-scale feature network to generate the text boundary predictions at different scales. Due to the unique elongated structure of the text, the algorithm uses conventional square receptive field convolution, which not only introduces background noise but also has a limited perception range when extracting text features. Therefore, the researchers have designed various convolutions with elongated receptive fields [17,24,25] and deformable convolution operations [26] to focus the convolution area on the text area and suppress background interference; and the transformer architecture [27,28] is introduced to enhance the long-range modeling ability of text detectors to compensate for the shortcomings of convolutional receptive fields.

In observing existing text feature extraction networks, it becomes apparent that many adopt the architecture of general object feature extraction networks. However, these networks lack the specific designs tailored for text features. While mature feature extraction networks yield effective results, they often compromise interpretability and make it challenging to design subsequent text feature adaptation components. To mitigate the impact of feature extraction networks on our approach, our designed algorithm architecture offers flexibility. It can seamlessly utilize both convolutional networks and transformer networks for learning text features. This adaptability ensures that our extraction approach remains unaffected by the limitations of specific feature extraction network architectures, providing a more versatile and effective solution.

2.2. Contour Representation for Scene Text Detector

Text instances have evolved from horizontally distributed ICDAR 2013 [1] to obliquely distributed ICDAR 2015 [2] and MSRA-TD500 [3], and then to curved distributed CTW1500 [4]. The form of text presentation is constantly becoming more diverse, and the ability of detection algorithms to express text contours is constantly improving. From the early expression of rectangular boxes [17,29] and inclined rectangular boxes [18,19,24], to the current expression of arbitrary shapes [8,10,13,23,30–32]. The contour point sequence estimation method based on the Cartesian coordinate system [33,34] and the arbitrary shape expression method based on polar coordinates [30] not only require later conversion processing, but are also inaccurate for text positioning with large-aspect ratio differences and height curvature. FCENet [10] first converts the problem to the Fourier domain for solution and then uses Fourier signals to fit an arbitrarily shaped text. Many algorithms [8,13,27,31,32,35] use pixel-level learning-based mask layer representation methods to determine the boundaries of text through geometric calculations. Our method follows the commonly used pixel-level learning approach to learn text regions in order to achieve the representation of any text shape.

While existing text detection methods demonstrate commendable performance in feature learning and contour representation, a notable drawback is the need for tailored training and tuning for distinct datasets. It not only escalates training costs but also hinders the application of these algorithms across various scenarios. To tackle this challenge, we

introduce a style-aware learning method. This innovative method allows for the simultaneous learning of text instances from various scenes and shapes, mitigating the style differences and consequently enhancing the overall performance of our proposed method.

3. Style-Aware Learning Network

3.1. Motivation

To better understand the proposed style-aware learning network and the corresponding training strategy, we introduce them in an experiment-driven way. Firstly, we analyze the shortcomings of existing strategies and summarize these shortcomings to refine the feasibility of optimization strategies, and then verify the existence of feasibility through experiments, providing a theoretical and experimental basis for us to propose a novel training strategy.

3.1.1. Analysis Settings

The combined dataset used for all experiments consists of four datasets: ICDAR 2013 [1], ICDAR 2015 [2], MSRA-TD500 [3], and CTW1500 [4]. Since the number of training images varies from dataset to dataset, we expand each dataset by copying training images, so that each dataset has approximately 1000 training images to balance the influence of different datasets at the training stage.

3.1.2. Existing Training Strategies

Existing text detection methods consist of two kinds of training strategies, such as directly straining on a combined dataset, and pre-training on a combined dataset and then fine-tuning on a certain dataset. As shown in Table 1, we analyze the two existing training strategies through experiment results.

Table 1. Results of the different training strategies. CD denotes the combined dataset. Pre-training denotes the initializing of network parameters with the training results of the network on combined datasets. TC denotes the training time cost of the training strategies, which is a relative value. TD denotes the training datasets. IC13, IC15, MSRA, and CTW denote ICDAR 2013, ICDAR 2015, MSRA-TD500, and CTW1500, respectively. We use **bold** for the top three results for each column in the table.

Methods	TD	Pre-Training	TC	IC13 (%)	IC15 (%)	MSRA (%)	CTW (%)	Mean F-Measure (%)
Baseline	CD		1.00	63.91	74.47	**76.78**	78.10	**73.51**
	IC13	✓	1.22	**77.94**	42.98	30.89	36.87	49.15
	IC15	✓	1.27	**76.71**	**80.09**	37.30	42.11	59.37
	MSRA	✓	1.23	48.28	43.78	**76.45**	73.72	61.66
	CTW	✓	1.28	50.42	51.77	74.91	**81.90**	**65.21**
	IC13		0.22	**77.02**	32.63	23.52	32.32	42.61
	IC15		0.27	75.98	**79.82**	29.31	38.21	56.08
	MSRA		0.23	46.32	38.35	70.40	68.91	56.94
	CTW		0.28	47.16	49.44	72.63	**81.78**	63.07
SLNText	CD		1.03	75.06	**76.80**	**77.47**	**79.42**	**77.36**

The first training strategy is to pre-train the method on the combined dataset and fine-tune the method on a certain dataset. The comparative experiments are shown in Table 1. The first line represents the test results of the method, which are trained on the combined dataset. The four following lines represent the results of the method which are fine-tuned on the corresponding dataset. The experimental results show that the fine-tuning operation can greatly improve the performance of the method on a certain dataset. One exception is the MSRA-TD500 dataset on the which fine-tuning effects are close to those of pre-training. We analyze that the reason is that the text styles of some images in CTW1500 and MSRA-TD500 are similar, so the training of CTW1500 can also improve

the performance of the method on MSRA-TD500. Although fine-tuning has an obvious improvement effect on the fine-tuned dataset, it significantly affects the performance of the method on other datasets. Take the second line as an example: the method that is fine-tuned on ICDAR 2013 brings 30%, 45%, and 42% F-measure reductions for the other three datasets, respectively. Additionally, the generalization of the dataset also affected this result; for example, the method which is fine-tuned on CTW1500 brings 13%, 23%, and 2% F-measure reductions for the other three datasets, respectively. This demonstrates that the CTW1500 dataset is highly generalized because it contains images similar in style to ICDAR2013 and MSRA-TD500. In general, the fine-tuning operation has a negative effect on the mean F-measure of the combined dataset, and the worse the generalization of the dataset, the greater the negative effect of fine-tuning. Since the combined dataset contains all styles of text, it is the most generalized dataset, so the method's performance in the multi-style text detection scenarios will only become worse regardless of which dataset we choose to use for fine-tuning.

The second training strategy is to train directly on a combined dataset. The comparative experiments are also shown in Table 1. The first line is the result of this training strategy. Lines 6–9 represent the results of the method which are only trained on the corresponding dataset. The experiment results show that, even though the training on the combined dataset can improve the mean F-measure of the method by nearly 8%, the performance on each dataset is worse than the method that is trained on a single dataset. MARA-TD500 remains the exception and the reason can be concluded from the ninth line that the CTW1500 can provide more diverse training samples with a similar style to MSRA-TD500 so that the method which is trained on CTW1500 performs better on MSRA-TD500 than the method which is trained on MSRA-TD500. Except for MASR-TD500, the other experimental results show that training samples of different styles interfere with each other, and simply mixing training images with different text styles as a training dataset is not an optimal choice to solve the problem of multi-style text detection.

3.1.3. Feasibility Analysis

After analyzing existing training strategies, we conclude that, if we want the method to perform well in multi-style text scenes, then the method needs to train different styles of text detection in parallel as independently as possible without affecting each other. In this section, we analyze the feasibility conditions of the parallel independent training. There are two main findings about the feasibility conditions. One is that the images in the same dataset have certain similarities and the images between the different datasets are distinguishable, so the method can accurately classify the images to the corresponding datasets through the feature extraction network. The other one is that a shared feature extraction network and a proprietary detection head can greatly reduce the influence between the different styles of text.

We design a simple classification network with ResNet50 as the backbone to classify the styles of the images, and the experimental results of this classification task are shown in Table 2. Each line in Table 2 represents the proportion of images from a particular dataset to be classified into four datasets. The classification network generates feature vectors for the multi-scale feature maps by the feature extraction network through average pooling and then classifies them through the fully connected network. The experimental results show that, with the exception of the ICDAR 2013 dataset, the probability that images from all datasets are correctly classified is more than 90%. It can be seen from the results that the ICDAR 2013 dataset is somewhat similar to the MSRA-TD500 and CTW1500 datasets, while the ICDAR 2015 dataset is different from the other three datasets, and the CTW1500 dataset is general. Therefore, nearly 19% of images from the ICDAR 2013 dataset are misclassified into the CTW1500 dataset. This is due to the similarity between datasets and does not mean that the network cannot distinguish between different styles of text.

The experimental results of a shared feature extraction network and proprietary detection heads are shown in the Table 3. The first line denotes the results of the baseline

directly trained on the combined dataset. The second line denotes the baseline only trained on the corresponding single dataset. The third line denotes that we use the combined dataset to train the shared feature extraction network; then, we freeze the parameters of the feature extraction network and use specific datasets to fine-tune the proprietary detection heads. It can be seen that not only do shared features not degrade the performance of the algorithm much, but they can also improve performance on some datasets. Additionally, the proprietary detection heads according to the text style can effectively avoid the mutual influence of different styles of text in the training stage.

Table 2. The text style classification results of the testing images from different datasets with a well-trained classification network.

Test Datasets	IC13 (%)	IC15 (%)	MSRA (%)	CTW (%)
IC13	75.54	0.43	5.58	18.45
IC15	0.00	97.60	0.80	1.60
MSRA	1.50	0.00	92.50	6.00
CTW	4.20	0.20	1.00	94.60

Table 3. Influence of shared feature extraction network (FEN) and shared detection head (SDH). MF denotes mean F-measure.

FEN	SDH	IC13 (%)	IC15 (%)	MSRA (%)	CTW (%)	MF (%)
✓	✓	63.91	74.47	76.78	78.10	73.51
		77.02	79.82	70.40	81.78	77.26
✓		81.44	78.11	78.24	81.32	79.78

3.2. Network Architecture

According to the previous analysis, we find that multi-style text detection can be realized by classifying the text style of the image, sharing a feature extraction network, and using the detection head dedicated to the dataset. Therefore, we develop a style-robust text detection network based on style-aware learning. As shown in Figure 1, this includes two momentous components: a style-aware head for distinguishing text styles, and a dynamic selection head for accurately predicting the text instances with different styles. Additionally, an efficient training strategy is designed to help SLNText obtain a satisfied performance on the different datasets while keeping nearly the same training and inference time cost as the baseline.

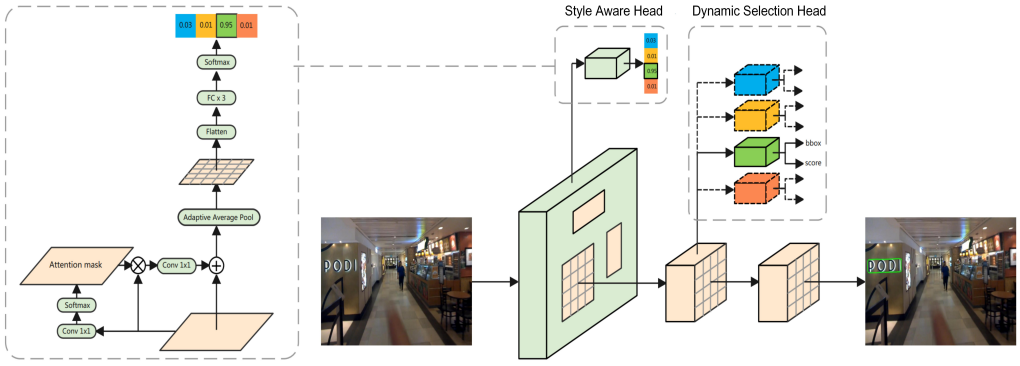

Figure 1. Architecture of the proposed style-aware learning network, which consists of the feature extraction network, style aware head, dynamic selection head, and instance segmentation head.

3.2.1. Architecture Design

The SLNText is proposed based on Mask R-CNN, and the architecture of the method is shown in Figure 1. The method can be divided into four modules: feature extraction network, style aware head, dynamic selection head, and instance segmentation head. The

feature extraction network and the instance segmentation head are the same as Mask R-CNN, we will describe the other two momentous modules in detail.

Style-aware head is added to classify the input image into different text styles and its architecture is shown in Figure 1. The feature maps output by the feature extraction network have five scales and the number of channels of the feature maps at each scale is 256. The feature maps at each scale go through a 1×1 convolution and the softmax operation to obtain an attention mask map, respectively. Finally, the attention mask map is multiplied by the input feature maps and we add the output to the input feature maps. Then, the average pooling is applied to obtain five vectors of length 256. These vectors are concatenated together and go through a fully connected three-layer network to obtain the final text style classification vector. After a softmax operation, the vector can represent the text style of the input image.

Dynamic selection head uses different detection heads to detect bounding boxes based on the classification result of the text style head. It consists of N detection heads and N means the number of text styles. Each detection head is the same as Mask R-CNN and an input image activates only one detection head, either during the training or testing stage.

3.2.2. Dynamic Training Strategy

To adapt to the proposed SLNText, we propose a dynamic training strategy (DTS), which can train our network well at the same training cost as the baseline algorithm. The pseudo-code of our proposed training strategy is in Algorithm 1. During training, we mix images from different datasets in a batch. Each image obtains a label indicating its dataset. Then, the batch goes through the network for feature extraction and alignment. We tag the features with the corresponding dataset label. These labeled features are sent to the right detection head and matched with their dataset label to calculate the loss. This process helps our algorithm learn from various datasets in one go, making it adaptable and effective across different scenes.

Algorithm 1: Dynamic training strategy

Input:
$\mathcal{I} \in R^{B\times 3\times H\times W}$ and $\hat{\mathcal{S}} \in \mathcal{C}^B$ are a batch of images and its corresponding text style labels.
$\hat{\mathcal{A}}$ is the annotations of original text detection tasks such as bounding boxes and text segmentation maps.
Parameter:
$\mathcal{C} = \{0, 1, 2, \cdots, C-1\}$ is the text style index set.
C is the number of text styles.
B is the training batch size.
Output:
$\mathcal{L}$ is the loss of a single iteration.

$\mathcal{M} = \text{FeatureExtractionNetwork}(\mathcal{I})$
$\mathcal{S} = \text{DynamicStyleHead}(\mathcal{M})$
$\mathcal{L}_{style} = \text{CrossEntropyLoss}(\mathcal{S}, \hat{\mathcal{S}})$
$bbox, score = \varnothing, \varnothing$
for *each* $c \in \mathcal{C}$ **do**
    m = ($\hat{\mathcal{S}}$ == c)
    $\mathcal{M}_i = \mathcal{M}[m]$
    $bbox_i, score_i = \text{DetectionHead}_i(\mathcal{M}_i)$
    $bbox \leftarrow bbox_i$
    $score \leftarrow score_i$
$\mathcal{B} = \text{Concatenate}(bbox)$
$\mathcal{S} = \text{Concatenate}(score)$

Algorithm 1: *Cont.*

$\mathcal{L}_{det} = \text{DetectionLoss}(\mathcal{M}, \mathcal{B}, \mathcal{S}, \hat{\mathcal{A}})$
$\mathcal{L}_{seg} = \text{SegmentationLoss}(\mathcal{M}, \mathcal{B}, \mathcal{S}, \hat{\mathcal{A}})$
$\mathcal{L} = \mathcal{L}_{style} + \mathcal{L}_{det} + \mathcal{L}_{seg}$
return $\mathcal{L}$

In the testing stage, the input image first passes through the text style head to identify the style category. Then, based on this style category, the corresponding detection head provides the classification confidence and offset regression for the proposals within the dynamic selection head. These refined proposals are projected onto feature maps, and their features are used as input for the instance segmentation head, ultimately generating the final detection result.

4. Experiments

To verify the effectiveness of the proposed method, we conducted quantitative experiments on the four datasets, consisting of ICDAR 2013 [1], ICDAR 2015 [2], MSRA-TD500 [3], and CTW1500 [4]. At the training stage, the four datasets were mixed and input into the proposed method with their own dataset category labels.

4.1. Implementation Details

4.1.1. Datasets

ICDAR 2013 is a dataset focused on horizontal text in the scene. It contains 229 training images and 233 testing images with only horizontal texts. ICDAR 2015 contains 1000 training images and 500 testing images. The dataset contains multi-oriented scene text for incidental scene text detection. The scene text regions are annotated by 4 vertices of word-level irregular quadrangles. MSRA-TD500 contains 300 training images and 200 testing images. The dataset contains multi-oriented scene text in Chinese and English, with resolutions ranging from 1296×864 to 1920×1280. The text instances are labeled at the text-line level. CTW1500 is a dataset which focuses on the curved text. It consists of 1000 training images and 500 testing images. The text instances are labeled at the text-line level.

4.1.2. Training

We use Mask R-CNN [36] implemented on the opening mmocr [37] as the baseline for all experiments because many scene-text detection methods [38–40] based on Mask R-CNN have good performance. We use the ImageNet [41] pre-trained on ResNet-50 [20] with a 5-level feature pyramid structure [21] as the backbone. The whole network is trained using the stochastic gradient descent (SGD) algorithm for 160 epochs. We set the initial learning rate as 0.05 and decay it by 0.1 at epochs 80 and 130, respectively. Fine-tuning in some experiments means that, after the joint training for 160 epochs, another 160 epochs are individually trained on a certain dataset.

4.1.3. Inference

Considering that the testing image has no text style prior information, and all testing images are set to have the same test size. We resize the long side of the testing image to 1080 and keep the aspect ratio unchanged. The evaluation method we use is the official evaluation protocol of ICDAR 2015. We use the evaluation method to calculate the F-measure for the four datasets, respectively, and calculate their mean values as the mean performance to measure the training strategy.

4.2. Results

4.2.1. Comparisons with Previous Methods

The experiment results of the proposed SLNText on the combined dataset and the baseline Mask R-CNN trained on the combined dataset and single dataset are shown

in Table 4. The first line represents the result of Mask R-CNN which is directly trained on the combined dataset consisting of the four datasets. The second line represents the result of Mask R-CNN which is pre-trained on the combined dataset and fine-tuned on the corresponding dataset. The third line represents the result of the proposed method which is directly trained on the combined dataset.

Table 4. Results of the proposed method and the Mask R-CNN which is trained on the combined dataset and on each corresponding dataset with the same test size for each dataset. TC denotes the training time cost of the training strategies, which is a relative value. TD denotes the training datasets. CD denotes the combined of all datasets. SD denotes the corresponding single dataset.

Methods	Pre-Training	TC	TD	IC13 (%)	IC15 (%)	MSRA (%)	CTW (%)	Mean F-Measure (%)
Mask R-CNN		1.00	CD	63.91	74.47	76.78	78.10	73.51
	✓	2.00	SD	77.94	80.09	76.45	81.90	–
Proposed method		1.03	CD	75.06	76.80	77.47	79.42	77.36

According to the results of the first line and the last line, the proposed method outperforms the baseline on all datasets. The proposed method is 3.85% higher than the baseline which is trained on the combined dataset on the mean F-measure. It can be seen from the results that the proposed method can effectively avoid the interaction of different styles of text during the training stage through the dynamic selection of the detection heads.

The qualitative results of the proposed method and the baseline are shown in Figure 2. We leverage different datasets to denote distinct scene text styles. The ICDAR 2013 dataset, for instance, captures the English scene text in a horizontally focused shooting direction. In contrast, the ICDAR 2015 dataset showcases English scene text in an oblique direction resulting from random shooting. Moving to the MSRA-TD500 dataset consists of Chinese and English scene text with oblique and long objects captured by mixed shooting. Lastly, the CTW1500 dataset portrays English scene text arranged in a curved direction obtained by focused shooting. We visually demonstrate the superior detection performance of our algorithm across various scene types when compared to the baseline. This substantiates our method's efficacy in handling diverse text styles and shooting directions, showcasing its robustness in real-world scenarios.

Figure 2. Qualitative results of the proposed method and baseline on four datasets, i.e., ICDAR 2013 [1], ICDAR 2015 [2], MSRA-TD500 [3], and CTW1500 [4]. Top: results of the baseline trained on the combined dataset. Bottom: results of the proposed method.

The pre-training and fine-tuning training strategy should be the theoretical ceiling for improving the method performance in a single dataset due to the method learning more training samples through pre-training and reducing the negative influence caused by the mutual interference of training samples with different text styles by fine-tuning. According to the results of the last two lines, the performance of the proposed method is close to the performance of the baseline, which is pre-trained on the combined dataset and fine-tuned on the corresponding single dataset. Even the proposed method performs better on the MSRA-TD500 dataset than the fine-tuned model. We conclude that the reason is that the shared feature extraction network for the combined dataset can also improve the performance of the method on some datasets that do not have enough training samples.

4.2.2. Ablation Study for Text Style Label

Compared with the baseline, the proposed method adds additional text style information of images as network supervision at the training stage to make a fair comparison. We conduct an experiment to explore the impact of adding the style-aware head to the baseline. The experimental results are shown in Table 5. The experimental results show that adding the style-aware head alone does not bring much improvement to the baseline, so it indicates that the improvement brought by our proposed method is mainly due to the dynamic selection head for the detection head.

Table 5. Results of the baseline (Mask R-CNN) and the impact of the style aware head and dynamic selection head for each dataset.

Methods	Style Aware Head	Dynamic Selection Head	IC13 (%)	IC15 (%)	MSRA (%)	CTW (%)	Mean F-Measure (%)
Baseline			63.91	74.47	76.78	78.10	73.51
	✓		65.60	76.22	75.59	77.32	73.98
	✓	✓	75.06	76.80	77.47	79.42	77.36

4.2.3. Ablation Study for Training Strategy

To verify the effectiveness of the proposed dynamic training strategy (DTS). We compared the proposed method with the mask R-CNN which is trained on the combined dataset, which we named CT. For the proposed method, if DTS is not used, we first pre-train the detection network and the proposed text style head on the combined dataset. Then, we fix the feature extraction network and the text style head and finetune the dynamic selection head on the corresponding dataset. As shown in Table 6, the proposed method without DTS outperforms the baseline by 3.94% in mean F-measure, but the training time cost is 1.48 times the baseline. The addition of DTS reduces the mean F-measure by 0.09% but reduces the training time cost by a large margin. With DTS, the proposed method can keep nearly the same training cost as the baseline but outperforms the baseline by 3.85% in mean F-measure.

Table 6. The impact of the dynamic training strategy (DTS) in the proposed method. TS denotes the training strategy. TC denotes the training time cost, which is a relative value.

Methods	TS	TC	IC13 (%)	IC15 (%)	MSRA (%)	CTW (%)	Mean F-Measure (%)
Baseline	CT	1.00	63.91	74.47	76.78	78.10	73.51
SLNText	(w\o) DTS	1.48	75.34	77.04	77.63	79.78	77.45
SLNText	DTS	1.03	75.06	76.80	77.47	79.42	77.36

4.2.4. Ablation Study for the Testing Size

To maintain the consistency of different datasets, we use the same test size for all test images, and the long side is set to 1080. However, this test size does not fit all datasets, leading to the baseline and proposed methods performing poorly on some datasets. To demonstrate the improved performance of our proposed algorithm on the optimal test size, we test the baseline and proposed method, respectively, with the best test size of each dataset. The long side for ICDAR 2013, ICDAR 2015, MSRA-TD500, and CTW1500 are 1080, 2200, 640, and 960, respectively. The experimental results are shown in Table 7. It can be seen that the proposed method still undergoes a significant improvement in mean performance with the best test size for each dataset. In particular, on the ICDAR 2015 dataset, our proposed method has basically the same effect as the baseline which is fine-tuned on the ICDAR 2015. This shows that, under the condition of the appropriate test size, the shared feature extraction network can also improve the performance of the algorithm on the ICDAR 2015 dataset.

Table 7. Results of the proposed method and the baseline which is trained on the combined dataset and fine-tuned on each dataset with the best test size for each dataset.

Methods	Pre-Training	TC	TD	IC13 (%)	IC15 (%)	MSRA (%)	CTW (%)	Mean F-Measure (%)
Baseline		1.00	CD	63.91	83.88	77.47	79.80	76.50
	✓	2.00	SD	77.94	84.46	75.53	82.10	–
SLNText		1.03	CD	75.06	84.58	77.90	80.74	79.65

4.3. Discussion

The proposed SLNText method efficiently tackles the issue of prolonged training times associated with existing algorithms. These algorithms often demand multiple fine-tuning iterations and the creation of numerous models, especially when confronted with significant disparities in scene text styles. We skillfully design a style-aware head and a dynamic selection head, catering to the nuances of various text styles. To maintain a training time comparable to the baseline algorithm, we introduce a dynamic training strategy. Unlike existing algorithms, which make a one-time training run on a combined dataset followed by multiple fine-tuning runs on different datasets, our method streamlines the process by requiring only a single training run on a combined dataset. This results in a noteworthy reduction in the training time. In the inference stage, existing algorithms need to know the scene type of each dataset in advance, and then use specific fine-tuning models. Our method easily processes all the known scene types of datasets using a single model.

Despite its success in mitigating training time challenges for perceptible scene types within the combined dataset, our proposed method does come with a noteworthy limitation. The effectiveness of our method relies on the inclusion of perceptible scene types in the combined dataset. When confronted with text from unknown scene categories, the generalization ability of our method is challenged. It is important to note that our method, which is essentially an enumeration-type scene robust solution, lacks adaptive capabilities for unknown scenes. The inherent limitation lies in its dependency on predefined scene types, making it less suitable for scenarios where the scene categories are unpredictable or extend beyond the initially considered set. Future enhancements may focus on augmenting adaptability to unknown scenes to broaden the applicability of our method.

5. Conclusions

In this paper, we present a style-aware learning network for accurately predicting the scene texts with varied styles. The proposed SLNText can efficiently tackle the problem that current algorithms experience in obtaining high performance across varied text styles by utilizing a single set of parameters and a one-time training process across numerous datasets. Our method distinguishes the text styles via the style-aware head and matches

the optimized prediction head via the dynamic selection head. Both of them are algorithm-agnostic, allowing for smooth integration into other algorithms and thereby ensuring broad applicability.

In our future works, we plan to enhance the generalization capability of SLNText by incorporating a zero-shot learning strategy. This strategy aims to broaden the model's ability to handle new and unseen scenarios. Additionally, we aim to bolster its text perception capacity in noisy images by introducing anti-interference modules, allowing the algorithm to detect challenging visual scene texts robustly. Another key focus is to make SLNText an end-to-end solution for text detection and recognition by integrating specialized recognizers into the system. These efforts are pivotal to creating a more general solution for text perceptions in varied scenarios.

Author Contributions: Project administration, Y.C. and F.Z.; methodology, Y.C. and F.Z.; validation, Y.C.; investigation, Y.C., F.Z. and R.Y.; writing—original draft, Y.C.; writing—review and editing, Y.C., F.Z. and R.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially supported by the Beijing University of Posts and Telecommunications Basic Research Fund with No. 2022RC12, the State Key Laboratory of Networking and Switching Technology with No. NST20220303, and the Beijing Natural Science Foundation with No. 4232025.

Data Availability Statement: The data can be downloaded from https://rrc.cvc.uab.es/?ch=4&com=introduction (accessed on 27 December 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; Bigorda, L.G.I.; Mestre, S.R.; Mas, J.; Mota, D.F.; Almazàn, J.A.; Heras, L.P.D. ICDAR 2013 robust reading competition. In Proceedings of the International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013.
2. Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V.R.; Lu, S.; et al. ICDAR 2015 competition on robust reading. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 1156–1160.
3. Yao, C.; Bai, X.; Liu, W.; Ma, Y.; Tu, Z. Detecting texts of arbitrary orientations in natural images. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1083–1090.
4. Yuliang, L.; Lianwen, J.; Shuaitao, Z.; Sheng, Z. Detecting curve text in the wild: New dataset and new solution. *arXiv* **2017**, arXiv:1712.02170.
5. Veit, A.; Matera, T.; Neumann, L.; Matas, J.; Belongie, S. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv* **2016**, arXiv:1601.07140.
6. Nayef, N.; Yin, F.; Bizid, I.; Choi, H.; Feng, Y.; Karatzas, D.; Luo, Z.; Pal, U.; Rigaud, C.; Chazalon, J.; et al. ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification-RRC-MLT. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 1454–1459.
7. Ch'ng, C.K.; Chan, C.S. Total-text: A comprehensive dataset for scene text detection and recognition. In Proceedings of the International Conference on Document Analysis and Recognition, Kyoto, Japan, 9–15 November 2017.
8. Liao, M.; Wan, Z.; Yao, C.; Chen, K.; Bai, X. Real-time scene text detection with differentiable binarization. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11474–11481.
9. He, M.; Liao, M.; Yang, Z.; Zhong, H.; Tang, J.; Cheng, W.; Yao, C.; Wang, Y.; Bai, X. MOST: A Multi-Oriented Scene Text Detector with Localization Refinement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8813–8822.
10. Zhu, Y.; Chen, J.; Liang, L.; Kuang, Z.; Jin, L.; Zhang, W. Fourier contour embedding for arbitrary-shaped text detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3123–3131.
11. Li, J.; Lin, Y.; Liu, R.; Ho, C.M.; Shi, H. RSCA: Real-time segmentation-based context-aware scene text detection. In Proceedings of the Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2349–2358.
12. Lin, Z.; Chen, Y.; Chen, P.; Chen, H.; Chen, F.; Ling, N. JMNET: Arbitrary-shaped scene text detection using multi-space perception. *Neurocomputing* **2022**, *513*, 261–272. [CrossRef]
13. Liao, M.; Zou, Z.; Wan, Z.; Yao, C.; Bai, X. Real-Time Scene Text Detection with Differentiable Binarization and Adaptive Scale Fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 919–931. [CrossRef] [PubMed]

14. Utterback, J.M.; Abernathy, W.J. A dynamic model of process and product innovation. *Omega* **1975**, *3*, 639–656. [CrossRef]
15. Sun, X.; Panda, R.; Chen, C.F.R.; Oliva, A.; Feris, R.; Saenko, K. Dynamic network quantization for efficient video inference. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 7375–7385.
16. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
17. Liao, M.; Shi, B.; Bai, X.; Wang, X.; Liu, W. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. In Proceedings of the AAAI, San Francisco, CA, USA, 4–9 February 2017.
18. Liao, M.; Zhu, Z.; Shi, B.; Xia, G.; Bai, X. Rotation-Sensitive Regression for Oriented Scene Text Detection. In Proceedings of the Computer Vision and Pattern Recognition, IEEE, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5909–5918.
19. Lyu, P.; Yao, C.; Wu, W.; Yan, S.; Bai, X. Multi-Oriented Scene Text Detection via Corner Localization and Region Segmentation. In Proceedings of the Computer Vision and Pattern Recognition, IEEE, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7553–7563.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
21. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
22. Wan, Q.; Ji, H.; Shen, L. Self-Attention Based Text Knowledge Mining for Text Detection. In Proceedings of the Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5983–5992.
23. Zhang, S.; Zhu, X.; Yang, C.; Wang, H.; Yin, X. Adaptive Boundary Proposal Network for Arbitrary Shape Text Detection. In Proceedings of the International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 1285–1294.
24. Cai, Y.; Wang, W.; Chen, Y.; Ye, Q. IOS-Net: An inside-to-outside supervision network for scale robust text detection in the wild. *Pattern Recognit.* **2020**, *103*, 107304. [CrossRef]
25. Du, B.; Ye, J.; Zhang, J.; Liu, J.; Tao, D. I3CL: Intra- and Inter-Instance Collaborative Learning for Arbitrary-Shaped Scene Text Detection. *Int. J. Comput. Vis.* **2022**, *130*, 1961–1977. [CrossRef]
26. Yang, Q.; Cheng, M.; Zhou, W.; Chen, Y.; Qiu, M.; Lin, W. IncepText: A New Inception-Text Module with Deformable PSROI Pooling for Multi-Oriented Scene Text Detection. In Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 1071–1077.
27. Ye, M.; Zhang, J.; Zhao, S.; Liu, J.; Du, B.; Tao, D. DPText-DETR: Towards Better Scene Text Detection with Dynamic Points in Transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 8–10 August 2023; pp. 3241–3249.
28. Huang, M.; Zhang, J.; Peng, D.; Lu, H.; Huang, C.; Liu, Y.; Bai, X.; Jin, L. ESTextSpotter: Towards Better Scene Text Spotting with Explicit Synergy in Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France 2–3 October 2023.
29. Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; Liang, J. East: An efficient and accurate scene text detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5551–5560.
30. Wang, F.; Chen, Y.; Wu, F.; Li, X. TextRay: Contour-based Geometric Modeling for Arbitrary-shaped Scene Text Detection. In Proceedings of the ACM International Conference on Multimedia, ACM, Seattle, WA, USA, 12–16 October 2020; pp. 111–119.
31. Wang, W.; Xie, E.; Li, X.; Hou, W.; Lu, T.; Yu, G.; Shao, S. Shape robust text detection with progressive scale expansion network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9336–9345.
32. Wang, W.; Xie, E.; Song, X.; Zang, Y.; Wang, W.; Lu, T.; Yu, G.; Shen, C. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8440–8449.
33. Liu, Y.; Chen, H.; Shen, C.; He, T.; Jin, L.; Wang, L. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–20 June 2020; pp. 9809–9818.
34. Liu, Y.; Shen, C.; Jin, L.; He, T.; Chen, P.; Liu, C.; Chen, H. ABCNet v2: Adaptive Bezier-Curve Network for Real-time End-to-end Text Spotting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 8048–8064. [CrossRef]
35. Zhang, S.X.; Zhu, X.; Hou, J.B.; Yang, C.; Yin, X.C. Kernel Proposal Network for Arbitrary Shape Text Detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *34*, 8731–8742. [CrossRef] [PubMed]
36. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
37. Kuang, Z.; Sun, H.; Li, Z.; Yue, X.; Lin, T.H.; Chen, J.; Wei, H.; Zhu, Y.; Gao, T.; Zhang, W.; et al. MMOCR: A Comprehensive Toolbox for Text Detection, Recognition and Understanding. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021.
38. Liao, M.; Pang, G.; Huang, J.; Hassner, T.; Bai, X. Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 706–722.

39. Wang, Y.; Xie, H.; Zha, Z.J.; Xing, M.; Fu, Z.; Zhang, Y. ContourNet: Taking a Further Step toward Accurate Arbitrary-shaped Scene Text Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11753–11762.
40. Xie, E.; Zang, Y.; Shao, S.; Yu, G.; Yao, C.; Li, G. Scene text detection with supervised pyramid context network. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
41. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.