Journal of Visual Language and Computing

journal homepage: www.ksiresearch.org/jvlc/

An Improved DBNet Model with Pyramid Pooling and Multi-Scale Enhancement for Small Text Object Detection

Zhu Yuan[a,1], Wuxuan Tang[b,*,2], Jin Yao[a,3] and Chengjun Liu[a,4]

[a]*Guangzhou Metro Design & Research Institute Co., Ltd., Guangzhou, China*

[b]*Institute of Intelligence Science and Technology, School of Computer Science and Software Engineering*

ARTICLE INFO

Submitted 7.31.2024
Revised 8.31.2024
Second Revision 10.31.2024
Accepted 11.15.2024

Keywords:
Small text object detection
DBnet
contextual information fusion
Dual attention

ABSTRACT

Detecting small text objects has been a key focus in object detection research due to their unique characteristics: small size, limited semantic information, susceptibility to interference in complex scenes, and tendency to be easily obscured, among others. At present, there are still two common issues in representative object detection models: First, small objects are easily interfered by the background or other objects, and second, multi-layer feature networks cause the loss of small object feature information. To address these challenges, this paper proposes an improved version of the DBNet model by introducing two modules: the contextual information fusion module SPP-CIF and the multi-scale feature enhancement module DA-MSFE. SPP-CIF fuses global and contextual information, by replacing the pooling layer of a pyramid with two sequentially concatenated dilated convolutions of small expansion rates, to encode semantic information of high-level features at multiple scales. DA-MSFE employs spatial attention and channel self-attention to select critical features at different scales and locations, and mines and exploits the correlations between channels to enhance and dynamically aggregate multi-scale features. Extensive experiments were conducted on the publicly available datasets MSRA-TD500 and ICDAR2015. The experimental results show that compared to the baseline model, the proposed model exhibits significantly superior performance in terms of the evaluation metrics.

1. Introduction

As an important carrier of information exchange and perception, text exists widely in daily life, such as advertising logos, promotional slogans, traffic signs, etc. Text object is quite unique, as it often located at the edges of images, far away, or in small fonts. Additionally, the factors such as varying aspect ratios, lack of clear closed contours, complex backgrounds, and lighting variations make small text difficult to detect. Consequently, small text object detection has become an important and challenging research topic.

In recent years, researchers have proposed various methods for text detection. Tian et al. [1] introduced a text detection framework with a vertical positioning mechanism called CTPN, which detects text lines within fine-grained text proposals in the convolutional feature map and extracts contextual information, effectively spotting deeply blurred text. Shi et al. [2] designed a directed text detection method, SegLink. It decomposes text into locally detectable segments and links, and enables full convolutional neural networks to perform dense detection at multiple scales through end-to-end training. Zhou et al. [3] proposed the EAST model, which employs multi-scale feature fusion to adaptively handle text of different sizes and predict words or text lines in arbitrary directions and quadrilaterals in complete images. Li et al. [4] introduced PSENet that utilizes a progressive scale expansion network to generate different scale kernels for each text instance, addressing the localization of arbitrarily shaped text. To alleviate the problem of poor detection of curved text, Long et al. [5] proposed a scene text representation, TextSnake, which better handles the detection of curved text. DBNet [6] improved the segmentation effects by using adaptive threshold maps

*Corresponding author
Email address: tangwx2001@hhu.edu.cn

DOI reference number: 10-18293/JVLC2024-N1-061

for training and introducing differentiable binarization to solve the gradient non-differentiability problem. More recently, the model Transformer has also been introduced into this field to tackle curved or polygonal scene text detection [7,8].

However, as one of the representative models for text detection, DBNet still suffer from two significant drawbacks:

Figure 1: An image where small texts are disturbed by the background.

Small text can easily be disturbed by the background, as shown in Figure 1. In natural scenes, the background of text images is complex and cluttered, and noise such as lighting interferes with the text detector, reducing the precision of the model and affecting the overall detection precision.

Figure 2: An image with text regions of different shapes.

Text regions with variable shapes are prone to omission. The aspect ratio and size of different text objects in the same image vary greatly, small-sized text is easily missed, and long text is difficult to detect completely. An image with text regions of different shapes is shown in Figure 2.

To address these issues, this paper proposes an improved model on the basis of DBNet, which introduces a Spatial Pyramid Pooling-based Context Information Fusion Module (SPP-CIF) and a Dual Attention-based Multi-Scale Feature Enhancement module (DA-MSFE) to the original model. The main contributions are summarized as follows:

SPP-CIF module: It replaces the pooling layer of the pyramid by applying a tandem dilated convolution in the last layer of the feature extraction network, to semantically encode the high-level features at multiple scales, and fuses the global and contextual information via the global pooling operation, which significantly reduce the interference of background noise.

DA-MSFE Module: Spatial attention weights the fused feature maps and generates feature map weights at different scales, and channel self-attention enhances inter-channel correlation through matrix manipulation and weight calculation. By selecting and aggregating features of different scales and locations, the omission rate of variable text region shapes is considerably reduced.

Experimental Validation: Experimental results on the publicly available datasets MSRA-TD500 and ICDAR2015 demonstrate that the improved model significantly outperforms the baseline model in precision, recall, and F-measure. Particularly, on the MSRA-TD500 Dataset, the improved model achieves an increase of 1.4%, 1.6%, and 1.5% in the metrics of precision, recall, and F-measure respectively.

The rest of this paper is organized as follows: Section 2 provides the overall structure of the improved model, Section 3 presents the Spatial Pyramid Pooling-based Context Information Fusion Module, Section 4 introduces the Dual Attention-based Multi-Scale Feature Enhancement Module, Section 5 conducts experiments and result analysis, and finally Section 6 concludes the paper.

2. Model Architecture

Figure 3 illustrates the architecture of the improved model, which differs from the original DBNet by incorporating two embedded modules SPP-CIF and DA-MSFE. The overall architecture consists of three parts: a feature extraction network, a feature fusion network, and a DBNet detection head. The feature extraction network consists of ResNet50 [9] and SPP-CIF. SPP-CIF is inserted to the last layer of the feature extraction network to fuse local contextual information and global feature information, obtaining the global contextual information of the feature map. The feature fusion network is composed of FPN [10] and DA-MSFE. DA-MSFE is inserted after FPN to enhance the fusion of features from four different scales. The loss function used in this model is consistent with that of DBNet.

3. Spatial Pyramid Pooling-based Context Information Fusion

As the depth of the network increases, the semantic information contained in the feature maps becomes richer. One can effectively capture the contextual information of the image by further extracting semantic information. PSPNet [11] utilized pyramid pooling to fuse features at different scales, thereby reducing the loss of contextual information in sub-regions. PANet [12] employed pyramid structure to extract and fuse contextual information, while also utilizing global pooling to obtain global information. Inspired by these

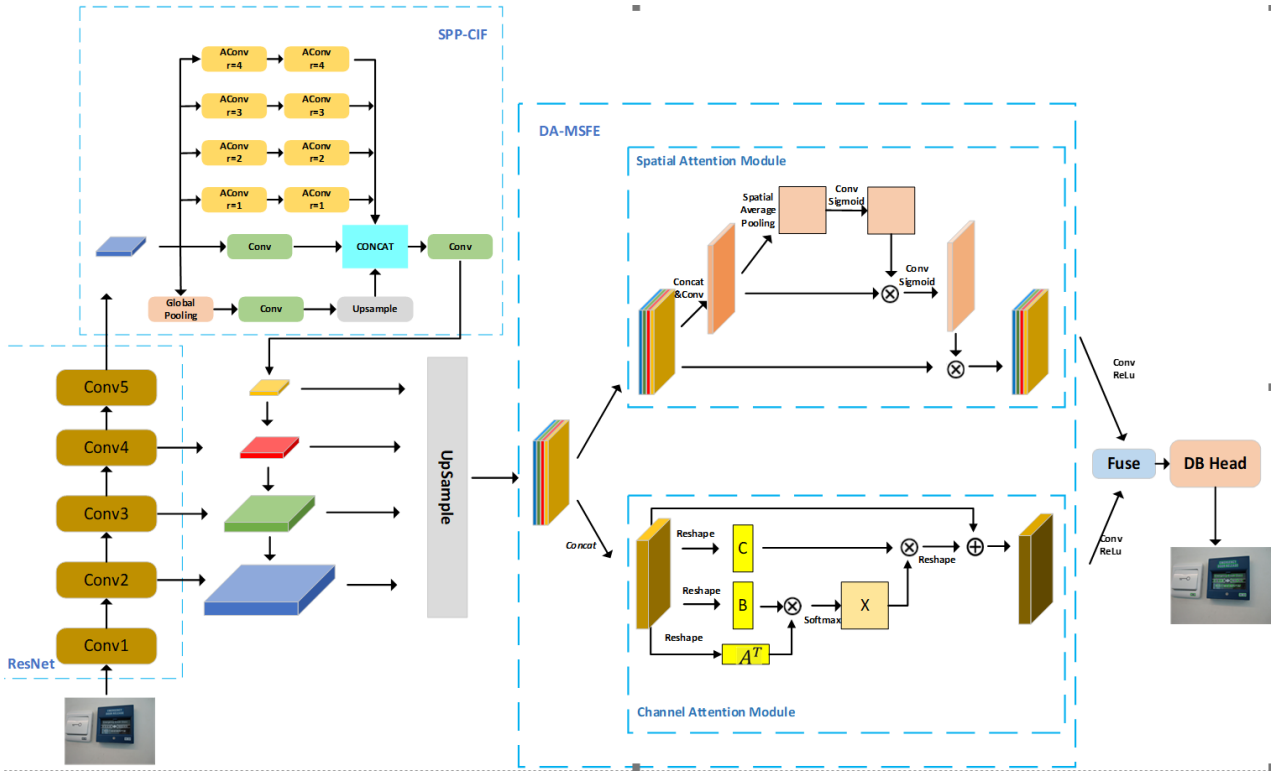

Figure 3: The architecture of the Improved DBNet Model with SPP-CIF and DA-MSFE

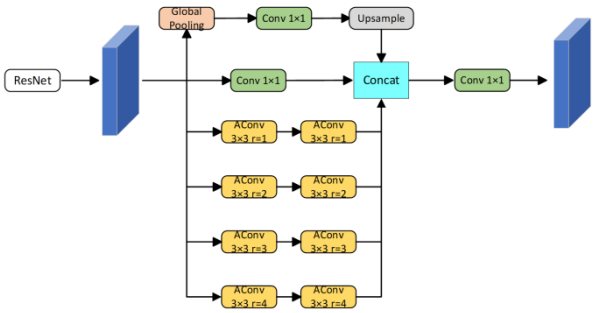

Figure 4: Spatial Pyramid Pooling-based Context Information Fusion Module SPP-CIF.

methods, this section proposes a Pyramid Pooling (SPP-CIF), which can extract information of high-level features from six different receptive fields.

The network structure of SPP-CIF is shown in Figure 4, and the module consists of a global pooling path, a convolutional path and a dilated convolutional path. The global pooling path is used to acquire global information to further improve the detection performance of the model. The convolutional path retains original feature information. The dilated convolution path employs four dilated convolutions with different dilation rates (r=1, 2, 3, 4) in parallel, increasing receptive fields and obtaining contextual information from diverse regions. Within the path, two Context Information Fusion Module based on Spatial dilated convolutions with small dilation rates are sequentially placed to comprehensively extract contextual information from high-level features, particularly focusing on small objects and their surrounding context. The information collected from six different receptive fields are concatenated along the channel dimension and then convolved by one layer to attain the fused feature map.

Specifically, the input feature map F_{in} of SPP-CIF is the output of the last layer of the feature extraction network. F_{in} obtains F_{in} information with different receptive fields from three respective paths. The process is as follows:

(1)The input feature map $F_{in}(F_{in} \in \mathbb{R}^{C\times H\times W})$ is fed to the global pooling path where there is a global average pooling operation, obtaining the global feature descriptor $F_{avg}(F_{avg} \in \mathbb{R}^{C\times 1\times 1})$. Then, a convolution operation with kernel size 1×1 is performed to gain the global information $F_u(F_u \in \mathbb{R}^{\frac{C}{2}\times 1\times 1})$. Finally, the information is upsampled to achieve $F'_u(F'_u \in \mathbb{R}^{\frac{C}{2}\times H\times W})$. This subprocess can be formulated as:

$$F'_u = Up\left(Conv_{1\times 1}\left(AvgPool(F_{in})\right)\right) \quad (1)$$

where $AvgPool$ denotes global average pooling, and Up represents upsampling operation.

(2)The input feature map $F_{in}$is input to the convolution path with kernel size 1×1 to obtain the feature map $F'(F' \in \mathbb{R}^{\frac{C}{2}\times H\times W})$, preserving some of the original information of the feature map. The formula for this subprocess is as follows:

$$F' = Conv_{1\times 1}(F_{in}) \quad (2)$$

where $Conv_{1\times1}$ denotes convolution with kernel size 1×1.

(3)The input feature map F_{in} is fed into a parallel dilated convolution pyramid network. This network uses dilated convolutions with kernel size 3×3 and dilation rates r of 1, 2, 3, and 4, respectively. Concatenating dilated convolutions with smaller dilation rates allows for the extraction of context information from different regions and focuses on small objects and their surroundings. This subprocess is formalized as follows:

$$A_i' = AConv_{3\times3,i}\left(AConv_{3\times3,i}(F_{in})\right), i = 1,2,3,4 \quad (3)$$

where $AConv_{3\times3,i}$ denotes dilated convolution with kernel size 3×3 and dilation rate of i.

The feature map with six types of information, F_u', F', and A_i' where $(i = 1,2,3,4)$, are concatenated along the channel dimension and then fed into a convolution with kernel size 1×1 for further fusion. The output feature map F_{out} *(where* $(F_{out} \in \mathbb{R}^{C\times H\times W})$ integrates both global information extracted from high-level feature maps and contextual information. The subprocess is expressed as follows

$$F_{out} = Conv_{1\times1}(Concat(F_u', F', A_1', A_2', A_3', A_4')) \quad (4)$$

where $Concat$ denotes concatenation operation along the channel dimension.

4. Dual Attention-based Multi-Scale Feature Enhancement

The feature fusion network outputs feature maps of four different scales, with downsampling factors of 4x, 8x, 16x, and 32x, respectively. These scale feature maps are upsampled to 1/4 of the original image size, and then subjected to feature enhancement by DA-MSFE. These feature maps have varying degrees of importance at different scales, and even within the same scale, the importance varies. Attention mechanisms enable the model to focus more on object regions with valuable information, thereby improving the efficiency and generalization capability. Convolutional Block Attention Module [13] concatenates channel attention and spatial attention to focus on important object regions. Liao et al. [14] construct Adaptive Scale Fusion to learn weights at different scales and spatial locations in the spatial dimension, achieving scale-robust feature fusion. Fu et al.[15] introduce self-attention to assign weights to each pixel in terms of the relationship between input data, thereby capturing dependencies between different positions in the sequence. Inspired by these methods, this section proposes the Dual Attention-based Multi-Scale Feature Enhancement Module (DA-MSFE) to enhance multi-scale features.

The Dual Attention-based Multi-Scale Feature Enhancement Module (DA-MSFE), as shown in Figure 5, takes as input the feature maps at four different scales output from the feature fusion network. The feature maps at different scales are upsampled to the same scale and fed into two sub-modules. Although upsampling brings these feature maps to the same scale, the features they contain come from differnt Operations such as average pooling, convolution, and activation are applied to gain the spatial attention weights of the fused feature map. These spatial attention weights are then employed to weight the fused feature map. Furthermore, convolution and activation operations are applied to the enhanced fused feature map to obtain spatial attention weights for the corresponding four scale feature maps.

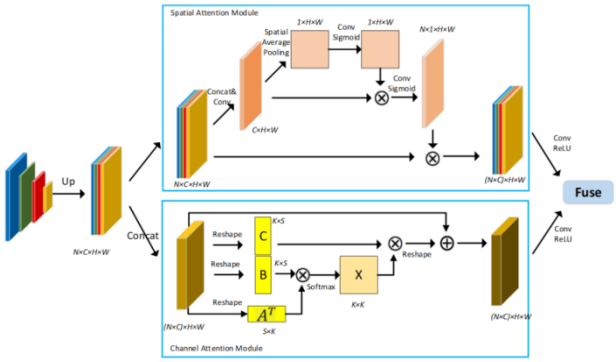

Figure 5: Dual Attention-based Multi-Scale Feature Enhancement.

Finally, these weights are utilized to weight the input four feature maps respectively. In the channel self-attention module, by reshaping the feature maps and performing matrix multiplication and weighting operations, each channel is assigned a weight that measures its relevance to other channels. The global information constructed from all channel weights is leveraged to enhance multi-scale features. The two sets of enhanced feature maps are convolved separately, added, and then fused by passing through another convolutional layer, to obtain the enhanced feature map.

The upsampling operation for DA-MSFE is formulated as follows:

$$\mathrm{F_{in}^{i}} = \mathrm{UpSample}(\mathrm{p_{i+1}}), \mathrm{i} = 1,2,3,4 \quad (5)$$

where $p_{i+1}(i = 1,2,3,4)$ represents the feature map with a scale of $\frac{1}{2^{i+1}}$ of the original image, $UpSample$ denotes upsampling, i.e., nearest-neighbor interpolation.

The operation of fusing the two enhanced feature maps in DA-MSFE is formalized as follows:

$$F_{out} = Conv_{3\times3,ReLu}\left(Conv_{3\times3,ReLu}(F_{outs}) + Conv_{3\times3,ReLu}(F_{outc})\right) \quad (6)$$

where F_{outs} is the feature map output from the spatial attention module, F_{outs} is the feature map output from the channel self-attention module, and $Conv_{3\times3,ReLu}$ represents the convolution with kernel size 3×3 and $ReLU$ activation.

The following subsections elaborate on the spatial attention and channel self-attention for multi-scale feature enhancement, respectively.

4.1 Generalities About the Model

For the feature maps of different scales upsampled to

the same size, the spatial attention module can capture feature information that the model focuses on from various perspectives and receptive fields. For instance, shallow, large-scale features can accommodate more detailed information and small text objects, while deep, small-scale features can capture richer high-level semantic information. To fuse and enhance features from different scales, instead of using simple summation, the spatial attention module DA-MSFE allows the model to autonomously choose important features from different scales and positions, dynamically aggregating features to achieve better integration.

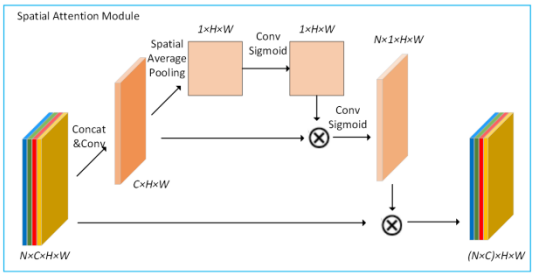

Figure 6: Spatial Attention Module.

The structure of the spatial attention module is shown in Figure 6, and its operation process is described below:

(1) Concatenate the four output feature maps $F_{in}^{i}(F_{in}^{i} \in \mathbb{R}^{C\times H\times W}, i = 1,2,3,4)$ to obtain F_{in}, then perform convolution with kernel size 3×3 on F_{in} to obtain the intermediate feature map $F_{in}^{i}(F_{in}^{i} \in \mathbb{R}^{C\times H\times W}, i = 1,2,3,4)$. This subprocess is formulated as follows:

$$F_{in}' = Conv_{3\times3}(Concat(F_{in}^{1}, F_{in}^{2}, F_{in}^{3}, F_{in}^{4})) \quad (7)$$

where $Concat$ denotes concatenation operation along the channel dimension.

(2) Perform global pooling on F_{in}' to obtain $F_{avg}(F_{avg} \in \mathbb{R}^{1\times H\times W})$, then apply a convolution with kernel size 3×3 to F_{avg} followed by a $Sigmoid$ function to obtain the descriptor $M_s(M_s \in \mathbb{R}^{1\times H\times W})$. Each position would bear a weight, allowing the model to learn the importance of each position in the fused feature map. This subprocess is formulated as follows:

$$M_s = Sigmoid\left(Conv_{3\times3}(AvgPool(F_{in}'))\right) \quad (8)$$

(3) Multiply the spatial descriptor M_s with the feature map F_{in}, then apply convolution with kernel size 3×3 to the result followed by a $Sigmoid$ function to obtain the spatial attention $A_s(A_s \in \mathbb{R}^{N\times1\times H\times W}, N = 4)$. This subprocess is expressed as follows:

$$A_s = Sigmoid(Conv_{3\times3}(M_s \otimes F_{in}')) \quad (9)$$

(4) Split the attention weights A_s into four attention weights A_s^{i} corresponding to the four scale feature maps F_{in}^{i}, perform weighting operation for each scale, and then concatenate them along the channel dimension to obtain the weighted feature map $F_{outs} \in \mathbb{R}^{(N\times C)\times H\times W}$. This subprocess is formulated as follows:

$$F_{outs} = Concat(A_s^{1} \otimes F_{in}^{1}, A_s^{2} \otimes F_{in}^{2}, A_s^{3} \otimes F_{in}^{3}, A_s^{4} \otimes F_{in}^{4}) \quad (10)$$

4.2 Channel Self-Attention for Multi-Scale Feature Enhancement

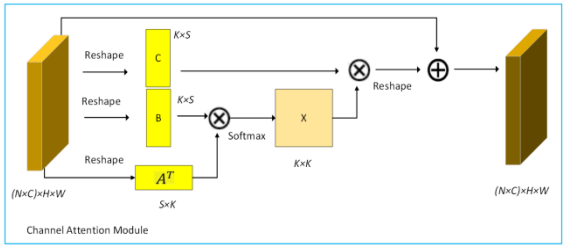

Figure 7: Channel Self-Attention Module.

The spatial attention module only considers the spatial information but neglects the channel information of different scale feature maps. In this section, we introduce the Channel Self-Attention Module to capture the correlation between the channels of these features.The global information consisting of correlations between channels is exploited to enhance the multiscale features.

Unlike traditional channel attention, the Channel Self-Attention Module does not utilize convolution to embed the feature maps. Instead, its implements feature embedding based on self-attention, which enables to fully explore the dependencies between all channels in the feature maps. The structure of the Channel Self-Attention is shown in Figure 7. Its operation process is as described below:

(1) Reshape the concatenated feature map $F_{in}(F_{in} \in \mathbb{R}^{(N\times C)\times H\times W})$ into three feature maps A, B, and C, where $\{A, B, C\} \in \mathbb{R}^{K\times S}, (K = N \times C, S = H \times W)$.

(2) Transpose feature map A and perform matrix multiplication between feature map B and A^T to obtain a $K \times K$ matrix X'. Apply the $Softmax$ function to X' to obtain the normalized channel attention weight matrix $X(X \in \mathbb{R}^{K\times K})$. $X_{i,j}$ represents the influence of the i channel on the j channel in the feature map, indicating the weight value. A higher weight value means a higher correlation between the two channels. This subprocess is formulated as follows:

$$X = Softmax(B \times A^T) \quad (11)$$

(3) Perform matrix multiplication between matrix X and feature map C to obtain the weighted feature map A_c. This operation is expressed as follows:

$$\mathrm{A_c} = \mathrm{X} \times \mathrm{C} \quad (12)$$

(4) Reshape the feature map A_c into $A_c^{T}(A_c^{T} \in \mathbb{R}^{(N\times C)\times H\times W})$, and add it to the input feature map F_{in} to obtain the final output feature map F_{outc}. This subprocess is formulated as follows:

$$F_{outc} = Reshape(A_c) + F \quad (13)$$

5. Experimental Results and Analysis

In this section we designed and conducted ablation experiments and comparative experiments to validate the effectiveness of the proposed model in detection of

small text objects. Below, we will introduce the datasets, evaluation metrics, implementation details, and analysis of the experimental results.

5.1 Datasets

Apparently, the types, scales, quantities, and qualities of objects form different datasets can all affect the learning performance of small object detection models. In the experiments, we utilized the following publicly available datasets:

(1) MSRA-TD500: It is published by Huazhong University of Science and Technology in 2012, containing this dataset contains 300 training images and 200 test images, with text boxes labelled as upper-left coordinates, width and height, and deflection angle.

(2) ICDAR2015: It is published by ICDAR in 2015, containing this dataset contains 1000 training images and 500 test images, with text boxes labelled as the four vertices of the polygon.

5.2 Evaluation Metrics

Since the text object is singular, three evaluation metrics are employed to assess the performance of the proposed model: Precision P, Recall R, and F-measure F. The formulas for calculating these metrics are as follows:

$$\begin{aligned} P &= \frac{TP}{TP+FP} \\ R &= \frac{TP}{TP+FN} \\ F &= 2 \times \frac{P \times R}{P+R} \end{aligned} \tag{14}$$

where TP refers to the number of positive samples that the model correctly predicts as positive, FP represents the number of negative samples that the model incorrectly predicts as positive, and FN denotes the number of positive samples that the model incorrectly predicts as negative. Precision measures the model's accuracy in predicting positives, indicating the proportion of predicted positive samples that are truly positive. Recall measures the model's coverage of positives, the proportion of positive samples that are successfully predicted by the model. F is the harmonic mean of precision and recall, used to balance precision and recall.

5.3 Experiment Setup and Implementation

We have set up two sets of experiments:

Ablation Experiments:This set of experiments aims to validate the performance of the two modules SPP-CIF and DA-MSFE in detecting small text objects. We use DBNet as the base model and train the following models: **Model 1**: the base model; **Model 2**: the base model with embedded SPP-CIF module; **Model 3**: the base model with the spatial attention part of DA-MSFE embedded; **Model 4**: the base model with the complete DA-MSFE module embedded; **Model 5**: the proposed model, DBNet-SD, i.e., the base model with embedded SPP-CIF and DA-MSFE modules.

Comparative Experiments: We compare the proposed model DBNet-SD with other commonly used text detection models on the ICDAR2015 and MSRA-TD500 datasets to validate its detection performance.

Due to the high randomness of the model initialization parameters, to accelerate model convergence, the backbone feature extraction network pretrained on the SynthText dataset is loaded, and then trained and tested on the datasets MSRA-TD500 and ICDAR2015. Since the dataset MSRA-TD500 has relatively few training images, 400 annotated images from HUST-TR400 are added to it to create a new training set, allowing the model to learn more features and achieve better detection performance.

During training, the input image size is set to 640×640, the number of training epochs is set to 1000, the batch size is set to 32, and the optimizer used is SGD with an initial learning rate of 0.001, following the Poly learning rate schedule. In addition to using random rotation, cropping, and flipping for image augmentation, we also conduct image scaling, skewing, and blurring for preprocessing, enhancing data diversity and further improving the generalization capability of the detection model.

The experiments were conducted on a platform featuring an Intel Gold 6146C CPU and an NVIDIA GeForce RTX 3090 GPU, and the operating system running on the platform is Linux and the CUDA version is 11.0.

5.4 Experimental Results and Analysis

(1)Ablation Experiments

Table 1: Results of ablation experiments on the dataset MSRA-TD500.

Number	Model	P(%)	R(%)	F(%)
1	**DBNet**	86.1	77.0	81.3
2	**DBNet + SPP-CIF**	87.2	78.0	82.3
3	**DBNet+DA-MSFE_SAM**	86.3	78.0	81.9
4	**DBNet + DA-MSFE**	86.9	78.3	82.4
5	**DBNet+SPP-CIF+DA-MSFE**	**87.5**	**78.6**	**82.8**

The ablation experimental results on the MSRA-TD500 dataset are shown in Table 1. The precision, recall, and F-measure of DBNet are 86.1%, 77.0%, and 81.3%, respectively.

The precision, recall, and F-measure of Model 2 are 87.2%, 78.0%, and 82.3%, respectively. Compared to Model 1, the values of the three evaluation metrics increase by 1.1%, 1.0%, and 1.0%, respectively. This indicates that the introduction of SPP-CIF can integrate contextual and global information and, suppress image background noise, thus enhancing text detection performance and reducing false positives.

The precision, recall, and F-measure of Model 3 are 86.3%, 78.0%, and 81.9%, respectively. Compared to Model 1, the values of the three evaluation metrics increase by 0.2%, 1.0%, and 0.6%, respectively, indicating that the spatial attention module can promote the model's focus on four different scale feature maps, achieving multi-scale feature enhancement.

The precision, recall, and F-measure of Model 4 are 86.9%, 78.3%, and 82.4%, respectively. Compared to Model 1, the values of three evaluation metrics improve by 0.8%, 1.3%, and 1.1%, with a comparatively noticeable improvement in recall. This indicates that the DA-MSFE module can enhance and integrate important information from multi-scale feature maps, and accurately locate text objects in feature maps of different scales, thus alleviating the issue of missed detections and improving detection performance.

The precision, recall, and F-measure of Model 5 are 87.5%, 78.6%, and 82.8%, respectively. Compared to Model 1, the values of three evaluation metrics improve by 1.%, 1.6%, and 1.5%, respectively. Moreover, Model 5 is also superior to Model 2 and Model 4 respectively. This indicates that the embedded modules SPP-CIF and DA-MSFE can work cooperatively in the base model to effectively alleviate the issues of small text object susceptible to background interference and feature loss.

(2) Comparative Experiments

The baseline models involved in the comparative experiments include RRD [16], TextBPN++ [17], PCR [18], FSG [19], EAST, SegLink, and DBNet. To validate the generalization performance of the proposed model DBNet-SD, training and testing of the involved models were conducted on the ICDAR2015 and MSRA-TD500 datasets, respectively.

Table 2: The experimental results on the dataset MSRA-TD500.

Number	Model	Backbone	P(%)	R(%)	F(%)
1	RRD	VGG-16	87.1	73.0	79.4
2	TextBPN++	ResNet-50	86.7	80.8	83.6
3	PCR	ResNet-50	86.5	77.1	81.5
4	FSG	ResNet-50	86.9	**81.0**	**83.8**
5	DBNet	ResNet-50	86.1	77.0	81.3
6	DBNet-SD	ResNet-50	**87.5**	78.6	81.3

The experimental results on the MSRA-TD500 dataset are shown in Table 2. The scores of precision, recall, and F-measure of DBNet-SD are 87.5%, 78.6%, and 82.8%, respectively. Compared to RRD, DBNet-SD shows increases in precision, recall, and F-measure by 0.4%, 5.6%, and 3.4%, respectively. Compared to TextBPN++, the values of precision increases by 0.8%. Compared to PCR, DBNet-SD manifests improvements in precision, recall, and F-measure by 1.0%, 1.5%, and 1.3%, respectively. Compared to FSG, the value of precision improves by 0.6%. The experimental results on the MSRA-TD500 dataset indicate that DBNet-SD has achieved competitive detection performance among the involved baseline models, especially in the metric of precision.

Apparently, the recall and F-measure of DBNet-SD are a bit lower than that of TextBPN++ and FSG on the dataset MSRA-TD500, respectively. This can be attributed to its focus on precision with a stricter detection criterion, leading to potentially miss some true text instances. In contrast, TextBPN++ and FSG might utilize a comparatively lower detection threshold, allowing them to discover a broader range of text instances.

Table 3: The experimental results on the dataset ICDAR2015

Number	Model	Backbone	P(%)	R(%)	F(%)
1	RRD	VGG-16	85.5	78.6	82.1
2	EAST	PVANet	81.1	72.9	76.8
3	SegLink	VGG-16	72.8	77.00	74.8
4	FSG	ResNet-50	87.8	**83.3**	**85.5**
5	DBNet	ResNet-50	88.0	82.1	84.9
6	DBNet-SD	ResNet-50	**88.6**	82.5	85.4

The experimental results on the ICDAR-2015 dataset are shown in Table 3. The scores of precision, recall, and F-measure of DBNet-SD are 88.6%, 82.5%, and 85.4%, respectively. Compared to RRD, DBNet-SD shows increases in precision, recall, and F-measure by 3.1%, 3.4%, and 3.3%, respectively. Compared to EAST, DBNet-SD manifests increases in precision, recall, and F-measure by 7.5%, 9.6%, and 8.6%, respectively. Compared to SegLink, DBNet-SD demonstrates improvements in precision, recall, and F-measure by 15.8%, 5.5%, and 10.6%, respectively. Compared to FSG, DBNet-SD improves by 0.8% in precision. The results suggest that the improved model DBNet-SD can achieve superior small text object detection performance among the baseline models in scenarios with complex backgrounds.

Similarly, the recall and F-measure of DBNet-SD are slightly lower than that of FSG, which can also be attributed to its focus on the metric of precision.

Figure 8 illustrates some of the detection results of the improved model DBNet-SD and the base model DBNet on the ICDAR2015 dataset. These images feature indoor scenes where text objects are blurred due to light pollution. In the first row of pictures, the text objects under the second icon are detected on the left side, whereas they are missed on the right side. In the second row, more blurred texts are detected on the left side but overlooked on the right side. In the third row, the word "SingTd" in the upper-right corner is found on the left side, while it is not detected on the right side. Therefore, it is evident that the improved model can effectively mitigate the interference from light and other noises, demonstrating better detection capability than the base model.

Figure 8: Examples of detection results on the dataset ICDAR2015. The left and right column corresponds to DBNet-SD and DBNet, respectively.

Figure 9: Examples of detection results on the data set MSRA-TD500. The left and right column corresponds to DBNet-SD and DBNet, respectively.

Figure 9 shows some of the detection results of the improved model DBNet-SD and the base model DBNet on the Figure 8: Examples of detection results on the dataset ICDAR2015. The left and right column corresponds to DBNet-SD and DBNet, respectively. MSRA-TD500 dataset. Long text objects are present in all images. In the first row of images, the left side detects the Chinese texts “Press Here” and “Emergency Break” in the images, while the right side does not. In the second row, the word “full” is detected on the left side, while it is not detected on the right side. In the third row, the Chinese word “corporation” is completely detected on the left side, while it is not detected on the right side. Obviously, it suggests that the improved model DBNet-SD can effectively mitigate the issue that the edges of long text objects are frequently undetected in the base model.

From Figure 8 and 9, it can be summarized that both the improved model and the base model exhibit a certain degree of text area miss-detection, but the improved model shows significant effectiveness in reducing miss-detections of small text objects compared to the base model, which implies that the embedded modules SPP-CIF and DA-MSFE can promote the performance of text object detection, especially for the detection of small text objects.

6. Conclusion

This paper has proposed an improved model DBNet-SD for small text object detection, which integrates two novel modules: SPP-CIF and DA-MSFE into the base model DBNet. SPP-CIF merges the global information and context information from high-level features to promote the capability of understanding the context of objects. DA-MSFE enhances the important regions of feature maps at four different scales, by using spatial attention to dynamically aggregate features and channel self-attention to fully explore the dependencies among all channels in the multi-scale feature maps. Extensive experiments were conducted on publicly available datasets MSRA-TD500 and ICDAR2015. The experimental results show that the improved model significantly outperforms the base model, thus alleviating the issues of background interference, missed detection, and inaccuracy in small text object detection.

In future work, we shall continue to optimize the model DBNet-SD by fine-tuning the network para-meters while pursuing the balance between the two evaluation metrics precision and recall.

References

[1] Dai, P., Zhang, S., Zhang, H., et al., 2021. Progressive contour regression for arbitrary-shape scene text detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE. pp. 7393-7402.

[2] Fu, J., Liu, J., Tian, H., et al., 2019. Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE. pp. 3146-3154.

[3] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 770-778.

[4] Li, H., Xiong, P., An, J., et al., 2018a. Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180.

[5] Li, X., Wang, W., Hou, W., et al., 2018b. Shape robust text detection with progressive scale expansion network. arXiv preprint arXiv:1806.02559.

[6] Liao, M., Wan, Z., Yao, C., et al., 2020. Real-time scene text detection with differentiable binarization, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI. pp. 11474-11481.

[7] Liao, M., Zhu, Z., Shi, B., et al., 2018. Rotation-sensitive regression for oriented scene text detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp.5909-5918.

[8] Liao, M., Zou, Z., Wan, Z., et al., 2022. Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence 45, 919-931.

[9] Lin, T.Y., Dollár, P., Girshick, R., et al., 2017. Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 2117‑2125.

[10] Long, S., Ruan, J., Zhang, W., et al., 2018. Textsnake: A flexible representation for detecting text of arbitrary shapes, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer.pp. 20‑36.

[11] Shi, B., Bai, X., Belongie, S., 2017. Detecting oriented text in natural images by linking segments, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 2550‑2558.

[12] Tang, J., Zhang, W., Liu, H., et al., 2022. Few could be better than all: Feature sampling and grouping for scene text detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE. pp. 4563‑4572.

[13] Tian, Z., Huang, W., He, T., et al., 2016. Detecting text in natural image with connectionist text proposal network, in: Computer Vision‑ECCV 2016: 14th European Conference, Springer. pp. 56‑72.

[14] Woo, S., Park, J., Lee, J.Y., et al., 2018. Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer. pp. 3‑19.

[15] Ye, M., Zhang, J., Zhao, S., et al., 2023. Dptext-detr: Towards better scene text detection with dynamic points in transformer, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI. pp.3241‑3249.

[16] Zhang, S., Yang, C., Zhu, X., et al., 2023. Arbitrary shape text detection via boundary transformer. IEEE Transactions on Multimedia.

[17] Zhang, X., Su, Y., Tripathi, S., et al., 2022. Text spotting transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE. pp. 9519‑9528.

[18] Zhao, H., Shi, J., Qi, X., et al., 2017. Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 2881‑2890.

[19] Zhou, X., Yao, C., Wen, H., et al., 2017. East: An efficient and accurate scene text detector, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 5551‑5560.