applied sciences

MDPI

*Article*

# An Improved Differentiable Binarization Network for Natural Scene Street Sign Text Detection

**Manhuai Lu [1,*], Yi Leng [2], Chin-Ling Chen [3,4,*] and Qiting Tang [5]**

[1] College of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Zhongshan Institute, Zhongshan 528402, China
[2] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[3] School of Information Engineering, Changchun Sci-Tech University, Changchun 130600, China
[4] Department of Computer Science and Information Engineering, Chaoyang University of Technology, Taichung 413310, Taiwan
[5] School of Electron and Information Engineering, University of Electronic Science and Technology of China, Zhongshan Institute, Zhongshan 528402, China
* Correspondence: lumanhuai@gmail.com (M.L.); clc@mail.cyut.edu.tw (C.-L.C.)

**Abstract:** The street sign text information from natural scenes usually exists in a complex background environment and is affected by natural light and artificial light. However, most of the current text detection algorithms do not effectively reduce the influence of light and do not make full use of the relationship between high-level semantic information and contextual semantic information in the feature extraction network when extracting features from images, and they are ineffective at detecting text in complex backgrounds. To solve these problems, we first propose a multi-channel MSER (Maximally Stable Extreme Regions) method to fully consider color information in text detection, which separates the text area in the image from the complex background, effectively reducing the influence of the complex background and light on street sign text detection. We also propose an enhanced feature pyramid network text detection method, which includes a feature pyramid route enhancement (FPRE) module and a high-level feature enhancement (HLFE) module. The two modules can make full use of the network's low-level and high-level semantic information to enhance the network's effectiveness in localizing text information and detecting text with different shapes, sizes, and inclined text. Experiments showed that the F-scores obtained by the method proposed in this paper on ICDAR 2015 (International Conference on Document Analysis and Recognition 2015) dataset, ICDAR2017-MLT (International Conference on Document Analysis and Recognition 2017-Competition on Multi-lingual scene text detection) dataset, and the Natural Scene Street Signs (NSSS) dataset constructed in this study are 89.5%, 84.5%, and 73.3%, respectively, which confirmed the performance advantage of the method proposed in street sign text detection.

**Keywords:** street sign text detection; maximally stable extremum region; differentiable binarization network; feature enhancement; natural scenes



**Citation:** Lu, M.; Leng, Y.; Chen, C.-L.; Tang, Q. An Improved Differentiable Binarization Network for Natural Scene Street Sign Text Detection. *Appl. Sci.* **2022**, *12*, 12120. https://doi.org/10.3390/app122312120

Academic Editor: Hui Yuan

Received: 1 November 2022
Accepted: 24 November 2022
Published: 27 November 2022



## 1. Introduction

Text detection technology in natural scenes is currently a popular research area in the field of image processing, and it has been widely used for street sign recognition, translation of scene images, and text recognition of license plates and billboards [1]. The detection and recognition of the text information of street signs in natural scenes have received widespread attention from scholars in various countries [2]. The current work [3] mainly focuses on the text detection of road signs in fields and on highways, whose backgrounds are mostly large areas of sky and highway. The single background makes it easy to detect text information. However, in crowded city streets, the text information of street signs usually exists in a complex background. Tall buildings, pedestrian vehicles, and many

other objects with similar appearances to street sign text are often mistaken as street sign text by the detection network, which reduces the accuracy of the text detection, Eliminating the complex background beyond the target to be detected is also a concern of scholars in various countries [4]. In addition, natural scene street sign text is significantly affected by factors of the natural environment, such as lighting, obstructions, and shooting angles, making text detection in street sign images very challenging.

With the advancement of deep learning text detection algorithms, the focus of text detection has shifted from horizontal scene text detection [5] to more challenging slanted text detection and arbitrary shape text detection [6]. Network frameworks such as Mask R-CNN [7] have achieved good results in scene text detection, but most Mask R-CNN-based methods use simple single-scale convolutional layers for stacking, which does not make full use of high-level semantic information for the detection of multi-scale and arbitrarily shaped text. In addition, some methods incorporate the attention mechanism in feature pyramid networks (FPNs) and replace ordinary convolution with inflated convolution [8], but they still do not make full use of the low-level information of the network. This approach ignores the importance of low-level information for smaller text and text edge detection. In summary, the current methods for detecting street sign text in natural scenes still have the following problems: (1) it is not possible to remove the influence of complex backgrounds on street sign text detection, and (2) the high-level semantic information of the feature extraction network and the contextual information in the network are not fully utilized. Therefore, this study aimed to obtain a natural scene text detection method that can effectively reduce the influence of complex backgrounds and make full use of the contextual information of the feature pyramid network to achieve the effective detection of street sign text.

Through observation, it was found that although the shape of the text changed due to inclination and bending, the relationship between the pixel values of the same text region did not change. We used the maximally stable extremum region (MSER) [9] method to pre-process the image, classify the text region and non-text region of the image, and remove a large number of non-text regions, thereby reducing the impact of the complex background on the detection of this text region. On this basis, a differentiable binarization network [10] (DBNet) was used as the base network. Its feature pyramid component was improved to make full use of the high-level semantic information of the feature extraction network and the contextual information in the network to enhance the capability of the DBNet for image feature extraction based on the feature that the DBNet can set the binarization threshold adaptively and simplify the post-processing. Finally, the text regions classified by the MSER method were put into the improved DBNet network for further detection.

The main contributions of this paper are as follows: (1) A multi-channel MSER method is proposed, which uses R, G, B, and S channels to extract the natural scene street sign text area, effectively reducing the impact of strong light and complex background on the street sign text extraction; (2) A feature pyramid route enhancement (FPRE) module was proposed to improve the feature extraction network of DBNet model and enhanced the transmission of semantic information at the lower layers of the network; (3) A high-level feature enhancement (HLFE) module was proposed to make full use of the high-level semantic information of the network; (4) This paper constructed a natural scene street sign (NSSS) dataset for natural scene street sign text detection and used this dataset to evaluate the effectiveness of the method proposed in this paper.

## 2. Related Work

In recent years, researchers in related fields have conducted a considerable amount of work, and the existing deep learning-based methods can be broadly classified into three categories: regression-based methods, segmentation-based methods, and methods based on a mixture of regression and segmentation.

Regression-based methods predict text regions through strategies such as convolution and linear regression. Naiemi et al. [11] introduced a pipeline based on a convolutional neural network to obtain more advanced visual features, and they proposed a new algorithm for encoding pixel values that highlighted the texture of characters. Liu et al. [12] proposed a grouped channel combination block to implement data-driven anchor design and adaptive anchor assignment, and they proposed a uniform loss weighting model to mitigate the inconsistency between classification scores and localization accuracy. Lu et al. [13] improve the shrinkage algorithm of the bounding box, making the model more accurate in predicting the short edges of the text area, and add Feature Enhancement Module (FEM) to increase the receptive field of the model and enhance its detection ability for long text areas. Wan et al. [14] used self-attention-based text knowledge mining (STKM) in the training model to induce the convolutional neural network (CNN) backbone to display the feature information ignored by the original pre-trained model, thus improving the detection performance of the backbone network. Although the regression-based approaches have achieved good results in the detection of horizontal text, they are less effective at the detection of curved and slanted text.

The segmentation-based approach treats text detection as a semantic segmentation problem and achieves text detection by segmenting irregularly shaped text regions. PixelLink [15] was the first to propose this idea and conduct related research. Pixels in the same text instance were first connected without regressing the text position and then text boxes were extracted directly from the segmentation results to achieve segmentation-based text detection. Zhu et al. [16] proposed a text component extraction network for text detection in arbitrary shape scenes is proposed, which can detect different text components through two parallel branches. These two branches are the Feature Redistribution Module (FRM) and an improved Transformer decoder, which generate accurate text components to detect text instances. Zhu et al. [17] proposed a Fourier contour embedding method, which predicted the Fourier vector of text instances in the Fourier domain and then reconstructed the text contour point sequence through the inverse Fourier transform in the image space domain. This approach could accurately approximate any closed shape. Hu et al. [18] proposed a text contour attention detector that could accurately locate the text of any shape in any direction. Qiao et al. [19] proposed a recursive segmentation framework that expanded the recursive path and refined the previous feature mapping into internal states to improve the segmentation quality. Cai et al. [20] proposed a text detector, which dynamically generates independent text instance perceptual convolution parameters for each text instance from multiple features, thus overcoming some insurmountable limitations of arbitrary text detection and effectively formulating text detection tasks for arbitrary shape scenes based on dynamic convolution.

The methods based on a mixture of regression and segmentation combine the features of regression and segmentation models to improve the performance of text detection. The EAST model proposed by Zhou et al. [21] could generate word- or line-level predictions directly from the complete image using a single neural network, simplifying the intermediate steps, and leading to a substantial improvement in the accuracy and precision of the model. Li et al. [22] proposed an origin-independent coordinate regression loss and text instance accuracy loss on a pixel-based text detector, which alleviated the impact of the target vertex ordering and predicted the location of text instances more accurately. Liu et al. [23] proposed a semi-supervised scene text detection framework (Semi Text) using a pre-trained supervised model and an unlabeled dataset to train a scene text detector that was both robust and accurate.

Although deep learning methods that can acquire high-level features through convolutional operations have achieved good results in recent years, these methods rely excessively on the adjustment of network parameters and lack flexibility. Some researchers still rely on traditional methods, such as MSER methods, for preliminary text region extraction from images. He et al. [24] developed a contrast-enhanced maximally stable extremum region algorithm (CE-MSER) and combined it with CNNs to increase the robustness of the detec-

tion network. Mittal et al. [25] used the characteristics of DCT, important information in the image is found by selecting multiple channels, and texture distribution is studied based on statistical measurement to extract features. Then a deep learning model is proposed to eliminate false positives and improve the performance of text detection. Hua et al. [26] used a combination of MSER and cloud of line distribution (COLD) approaches to extract candidate regions of image text, and the extracted features were then sent to the CNN for extraction. This method exhibited better detection effects under low light.

Although the above methods have made some progress in the field of text detection, they still have the following problems in the detection of severely slanted and different-sized texts in complex scenes.

(1) Constrained by the candidate frame, regression-based networks are less effective at detecting text with large tilt angles.

(2) Segmentation-based methods do not work well for detecting small text instances with low contrast and text instances with complex layouts in images.

(3) Most of the methods are not able to effectively remove the interference of complex backgrounds and strong light on text detection, resulting in false detection.

For street sign text detection in natural scenes, compared with existing methods, our proposed method was designed to eliminate the impact of complex backgrounds and illumination on detection, as well as the effects of different sizes and shapes of text regions in the images during detection. The detection process is shown in Figure 1. Although the text in natural scenes is greatly affected by the lighting and shooting angles, the pixel values and pixel relationships between text remain constant during character changes, so a multi-channel MSER method is proposed to preprocess the image. Compared with the traditional MSER method, the method in this paper uses multiple channels are used to extract the maximally stable extremal regions, reducing the impact of complex scenes and strong light on text area detection. In addition, an improved DBNet network is proposed to further detect the image of the text area preliminarily extracted by the multi-channel MSER method. Compared with the previous works, the newly added FPRE module and the HLFE module can make full use of the information of the feature extraction network, and improve the detection effect of the network on different shapes, sizes, and oblique texts.

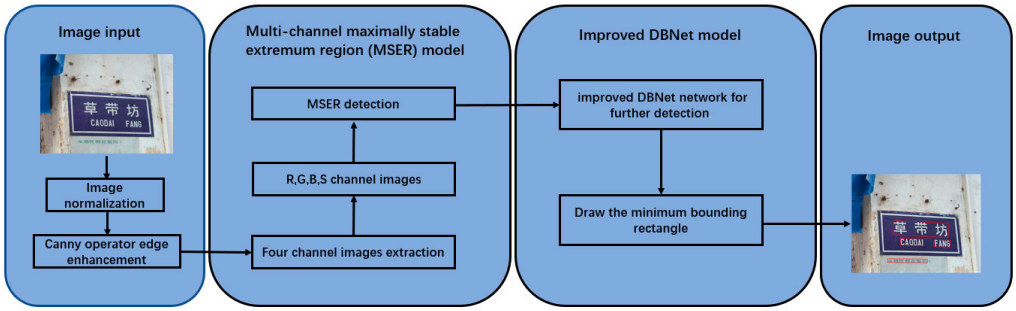


**Figure 1.** Overall flow chart of the method proposed in this paper.

## 3. Proposed Method

The proposed method includes a multi-channel maximally stable extremum region (MSER) model and an enhanced feature pyramid network text detection method.

### *3.1. Multi-Channel Maximally Stable Extremum Region (MSER) Model*

The pixels of the natural scene street signs have a certain relationship, which motivated us to explore the extraction of the maximally stable external regions in the natural scene street signs image. However, due to the influence of illumination in natural scenes, the brightness of different areas of the image is quite different, and only a single channel extracts a large and extremely stable area of natural scene images, which will cause some missed detections and false detections. The MSER method that is only under a single channel is difficult to adapt to the detection of street sign text affected by illumination and

complex background. Therefore, in addition to selecting the most commonly used RGB color channel of the image, we also studied the influence of the HIS color channel of the image on the extraction of the maximally stable external regions.

There are many color spaces for images, and the RGB color space is the most widely used color space. In addition, the HIS color space is often used in image processing. Therefore, in addition to selecting the most commonly used RGB color channel of the image, the experiment also studied the HIS. RGB and HIS color channel images are extracted, respectively, as shown in Figure 2. It can be found that the image in the S channel can effectively reduce the influence of strong light, and the text area is more obvious in the S color channel. The image under the H channel is greatly affected by the strong light, and the text area has been blended with the background under the influence of the strong light. The I-channel image can be replaced by different ratios of the R, G, and B channel information, as shown in (1). so the rendering effect of the image under the I channel is also similar to the effect of the RGB channel. Therefore, in the case where the maximally stable extremum region of the RGB three-channel has been extracted, it is of little significance to study the I-channel image. To explore the characteristics of the MSER method in different color channels, the Canny [27] operator was used to perform edge enhancement processing on the images of the RGB and HIS channels of the image, respectively, so that the text edges in the image become more obvious, thereby increasing the accuracy of the MSER algorithm, and then extracting the maximally stable extremum region from the image after edge enhancement, as shown in Figure 3. It can be found that the S channel can effectively extract the maximally stable extremum region under strong light, while the H channel is affected by strong light, and cannot effectively extract the maximally stable extremum region of text under strong light, and also introduces more non-text areas, which is not conducive to the further extraction of text areas using the deep learning framework. Therefore, it can be concluded that the S channel is very helpful for the initial extraction of the text area extracted by the MSER method, and the H and I channels are of little significance for the MSER method to extract the text area. Therefore, in the process of using the MSER method to extract the text area in multiple channels, we use $R$, $G$, $B$, and S color channels that contain important color information.

$$I = \frac{R + G + B}{3} \tag{1}$$

where the parameter $I$ is one of the channels of the HSI color space; $R$, $G$, and $B$ are the three channels of the RGB color space. According to the following equation [9] to determine whether they belong to the MSER:

$$\varphi(i) = \frac{|R_{i+\Delta} - R_{i-\Delta}|}{|R_i|} \tag{2}$$

where $R_i$ denotes a certain connected region when the threshold is $i$, $\varphi(i)$ denotes the area change rate of the region and $\Delta$ is the step size of the slight change of the grayscale threshold.

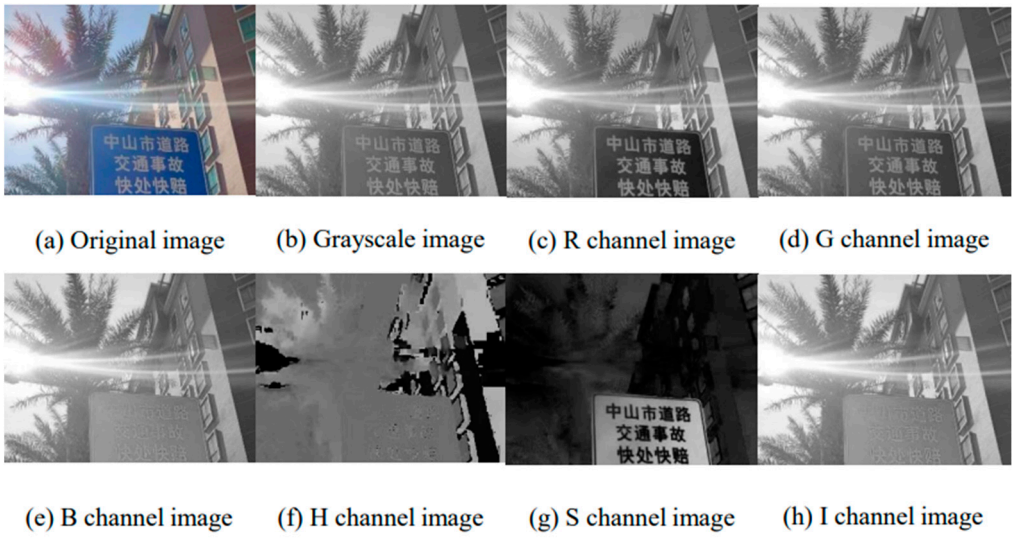


**Figure 2.** Image extraction results for different color channels.

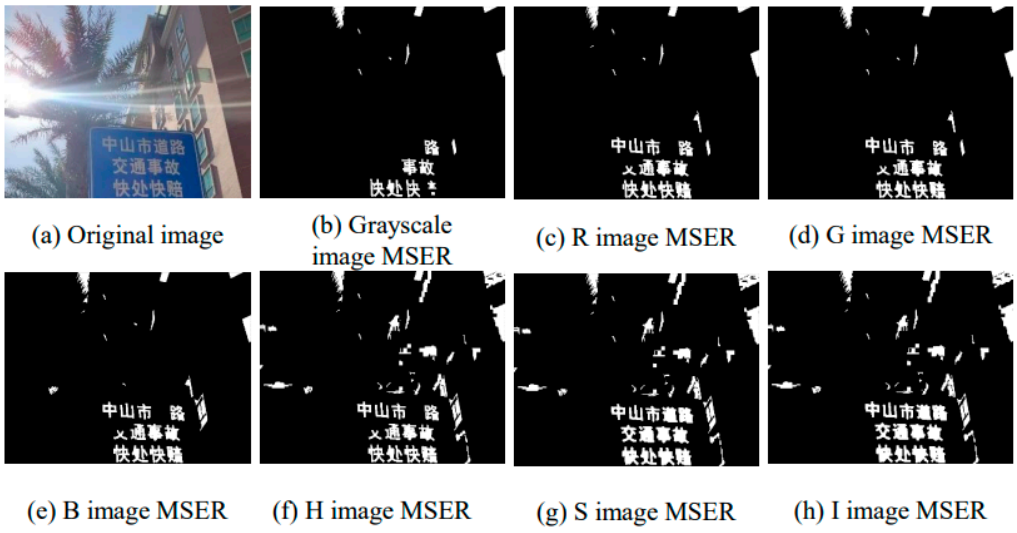


**Figure 3.** The MSER extraction results of different color channels.

To make the above statement more convincing, we compare the combination methods of different channels based on the NSSS dataset combined with the DBNet deep learning framework, and the experimental results are shown in Table 1. It can be seen from the table that the S channel does extract important information for the text area provider for the MSER method. However, H is greatly affected by strong light, and the I channel can be replaced by RGB channel information, which has little effect on the extraction of the MSER text area.

**Table 1.** NSSS dataset combined with DBNet text detection results in different channel combinations.

| Channel | Precision (%) | Recall (%) | F-Score (%) |
|---|---|---|---|
| RGB | 91.9 | 83.5 | 87.5 |
| RGBI | 91.8 | 83.76 | 87.6 |
| RGBH | 92.1 | 83.7 | 87.7 |
| RGBS | 92.3 | 84.2 | 88.1 |

### *3.2. Improved DBNet for Natural Scene Text Detection*

Although the multi-channel MSER model has removed most of the non-textual components of the image and reduced the complexity of the text detection problem, it is not sufficient to extract the textual regions of the street sign image, which motivated us to use an improved DBNet for further text detection. The DBNet algorithm is a segmentation-based network that can not only detect text in arbitrary directions but also use deformable convolution to enhance the detection of text with extreme aspect ratios. In addition, DBNet transforms the non-differentiable fixed binarization segmentation into an approximate differentiable binarization function, avoiding the problem of non-differentiable gradients, and it makes the final output image robust to the threshold value. While simplifying the post-processing, DBNet has better results on the lightweight backbone network, which accelerates the detection speed of the model. Therefore, we chose the DBNet model as the base network for street sign text detection in this paper, and the DBNet network structure is shown in Figure 4. However, the DBNet model adopts an FPN for feature fusion. When the FPN performs top-down feature fusion, the network expression ability is lost due to the mandatory reduction in feature channels, and there are missed and false detections for small-scale texts. In addition, the FPN is limited by the one-way information flow and the loss of low-level spatial information due to multiple sampling operations, which also have a certain impact on the accuracy of text detection.

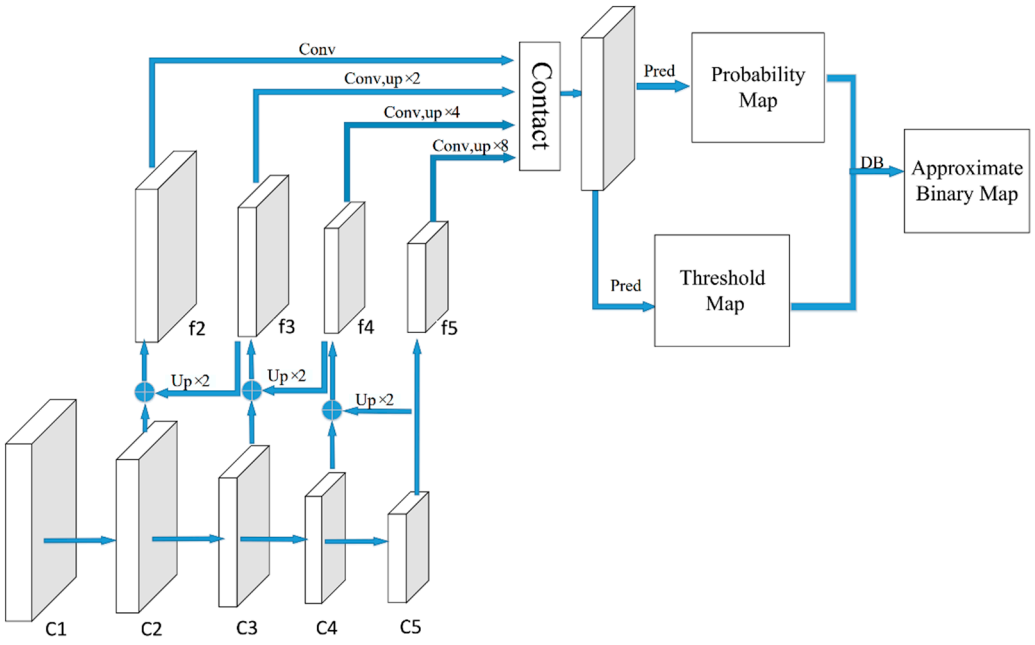


**Figure 4.** DBNet model network structure diagram.

In this study, the Resnet-18 [28] lightweight network was selected as the backbone network for feature extraction, which effectively reduced the computational complexity of the model. In order to be able to cope with text instances with extreme aspect ratios, all $3 \times 3$ convolutional layers were replaced with deformable convolution layers in the conv3, conv4, and conv5 stages of the feature extraction network of Resnet-18 [29] to improve the network's ability to focus on relevant image regions. At the same time, the HLFE and FPRE modules were introduced to enable the feature pyramid network of this model to fully utilize the high-level semantic information and contextual semantic information of the network. The structure diagram of the improved DBNet network is shown in Figure 5.

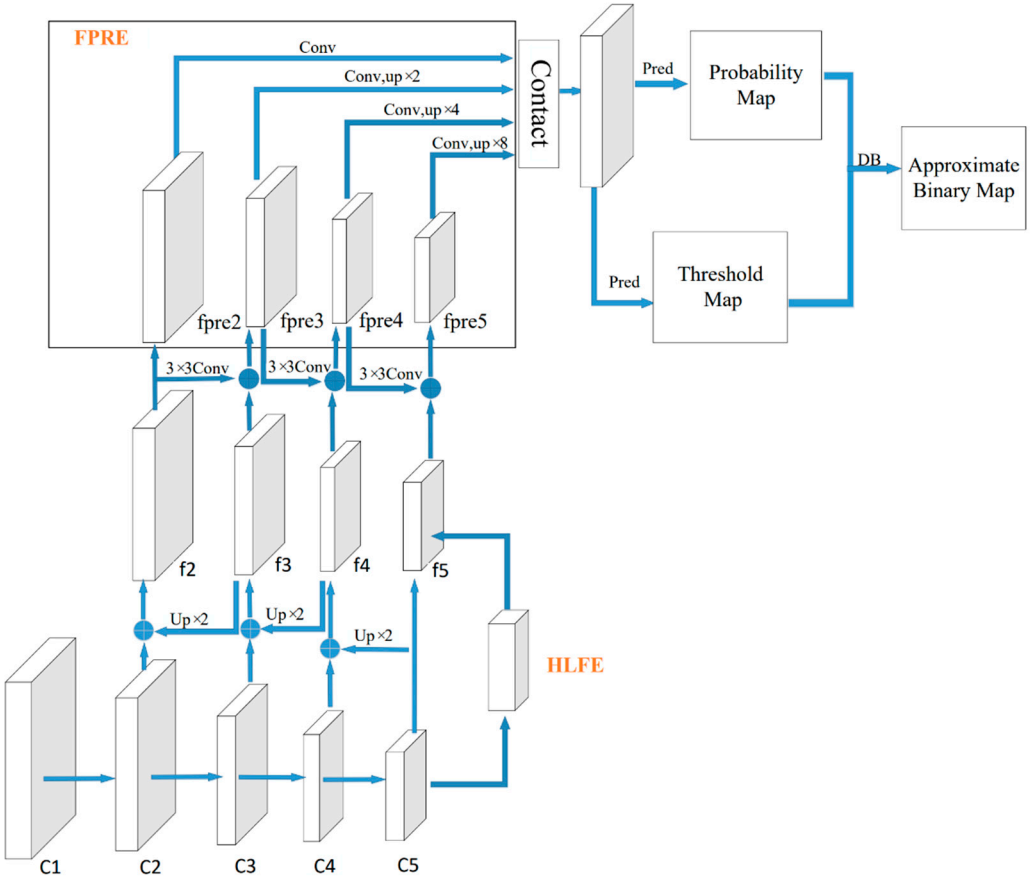


**Figure 5.** Network structure diagram of improved DBNet model.

#### 3.2.1. High-Level Feature Enhancement Model

For text detection networks, semantic contextual information plays a crucial role in scene text detection. To be able to detect both smaller and larger text in the same image and obtain good detection results, rich contextual information must be extracted from the image.

Resnet-18 first extracts feature maps {C2, C3, C4, C5} from the input images. C5, as the highest semantic feature map of the FPN, has many channels of information. However, when the FPN performs feature fusion from the top down, the number of channels of the top-level feature map must be reduced to the same number of channels of the bottom-level feature map to which it is fused to be able to fuse with the rest of the feature maps. Therefore, the number of channels is reduced in the process of downward propagation of the top-level feature map, resulting in the loss of semantic information in the network. To solve this problem, we proposed a feature enhancement module to enhance the high-level semantic information in this study, which reduced the loss of feature information and improved the detection effect of the net model.

As shown in Figure 6, first, the high-level semantic feature map *C5* was used as the input, and four feature maps of different scales were obtained after applying the following adaptive average pooling operation:

$$f_k = \alpha_k C5 : k = 1, 2, 3, 4 \tag{3}$$

where $k$ = 1, 2, 3, 4 denotes the four generated feature maps and $\alpha_k$ denotes the $\alpha_k$-fold downsampling operation on *f5*. In the experiments, $\alpha_1$, $\alpha_2$, $\alpha_3$, and $\alpha_4$ were set to 1, 0.5, 0.4, and 0.2, respectively.

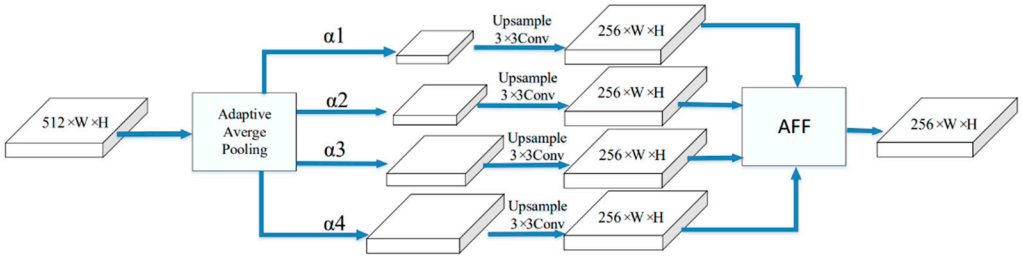


**Figure 6.** Schematic diagram of the network structure of the high-level feature enhancement (HLFE) module.

After downsampling, the channels of these four feature maps were changed to 256 dimensions through $1 \times 1$ convolutions, and the four feature maps were, respectively, normalized to a uniform scale by $\frac{1}{\alpha_k}$ upsampling through $3 \times 3$ depthwise separable convolution and linear interpolation upsampling methods. Then, the four feature maps were fused into a feature map of the same size as f5 through adaptive feature fusion (AFF), and finally, feature fusion was performed with f5.

The feature-enhanced network obtained different feature maps through adaptive pooling with a constant scale, and it obtained perceptual fields of different sizes after the $3 \times 3$ convolution layers, which increased the amount of information of the image context obtained by the top-level feature map. This facilitated text detection at different scales. In order to increase the convergence speed of the network, the parameters of the obtained feature maps were normalized by using BN layers, and they were finally passed to the rectified linear unite (ReLU) activation function and fused with f5 for features. In this way, the more contextual information of different spaces was incorporated into the feature map f5, so that the bottom-level feature map could obtain more semantic information propagated by the top-level feature map, thereby improving the text detection ability of the model.

#### 3.2.2. Feature Pyramid Route Enhancement Module

While strengthening the transmission of high-level semantic information of the model, the effect of low-level features of street sign images for text detection should not be underestimated. The low-level features contain very important information, such as the location and edge shape of the image. However, the low-level feature map must undergo multiple convolutions and sampling operations compared with the top-level feature map, which leads to the loss of useful information in the low-level feature map during the feature extraction process. In order to make full use of the low-level features and shorten the spatial information transmission distance, making the bottom-level information better propagate to the top level, we added a down-scale enhancement module based on the structure of the original FPN of DBNet. In the up-scale enhancement module of the original FPN, four feature maps, f1, f2, f3, and f4, with sizes of 1/4, 1/8, 1/16, and 1/32 of the input image, respectively, were obtained. In the down-scale enhancement module, the four feature maps obtained from the up-scale enhancement module were sampled and fused to obtain the feature maps fpre1, fpre2, fpre3, and fpre4 with sizes of 1/4, 1/8, 1/16, and 1/32 of the input image, respectively. With fpre3 as an example, f3 obtained a feature map of the same size as fpem2 through the upsampling operation and fused with fpem2 by element summation. The fused feature map was then convolved with a kernel size of 33 and a stride of 2 to generate fpre3 and then fpre4 was generated from fpre3 in the same way. The iterations continued until fpre5 was obtained. In the process of down-scale enhancement, all $3 \times 3$ convolutions applied depthwise separable convolution [30] instead of the original standard convolution, which greatly reduced the increase in the operational parameters and enabled the FPRE to obtain a larger perceptual field with a smaller increase in the number of parameters.

### 3.3. Loss Function

The improved DBNet loss function consisted of three components, which were the loss $L_s$ of the probability map, the loss $L_b$ of the binarized graph, and the loss $L_t$ of the threshold graph. The loss functions were as follows:

$$L = L_s + \beta_1 \times L_b + \beta_2 \times L_t \tag{4}$$

where $\beta_1$ and $\beta_2$ are the coefficients of the loss $L_b$ of the binarized graph and the loss $L_t$ of the threshold graph, which were, respectively set to 1.0 and 10 in this work. Furthermore, the loss $L_s$ of the probability graph and the loss $L_b$ of the binarized graph used the binary cross-entropy loss function to effectively solve the problem of unbalanced positive and negative samples, as follows:

$$L_s = L_b = \sum_{i \in R_d} y_i \log x_i + (1 - y_i) \log(1 - x_i) \tag{5}$$

where $x_i$ is the expected output of the sample, $y_i$ is the actual output of the sample, $R_d$ indicates the positive and negative samples that were sampled using OHEM (Online Hard Example Mining), and the sampling ratio was set to 1:3 based on the positive and negative sample regions of the image.

The loss $L_t$ of the threshold graph was used to calculate the sum of the distances between the predicted values and labels in the ground truth (Gd):

$$L_t = \sum_{i \in R_d} |y_i^* - x_i^*| \tag{6}$$

where $R_d$ is the index of the pixel values in the Gd, $x^*$ is the prediction result, and $y^*$ is the label of the threshold map.

## 4. Result and Discussion

### 4.1. Experimental Steps

To effectively detect street signs in natural scenes and test the effectiveness of our proposed method, we conducted the following experiments using the NSSS, ICDAR2015, and ICDAR2017-MLT datasets. (1) The images were directly input into the DBNet network for the experiments, and the results of this experiment were used as a reference to test the effectiveness of other methods. (2) The images were input into the DBNet network with a ratio-invariant feature enhancement module (HLFE). (3) The images were input into the DBNet network with a fused feature pyramid enhancement module (FPRE). (4) The images were input into the DBNet network with FPRE and HLFE modules. (5) The data set processed by multi-channel MSER was input into the modified DBNet network for the experiments to test the effectiveness of the MSER method. The overall flowchart of the experiments in this paper is shown in Figure 7, Y means to use this method for subsequent text detection experiments, and N means not to use this method for text detection experiments.

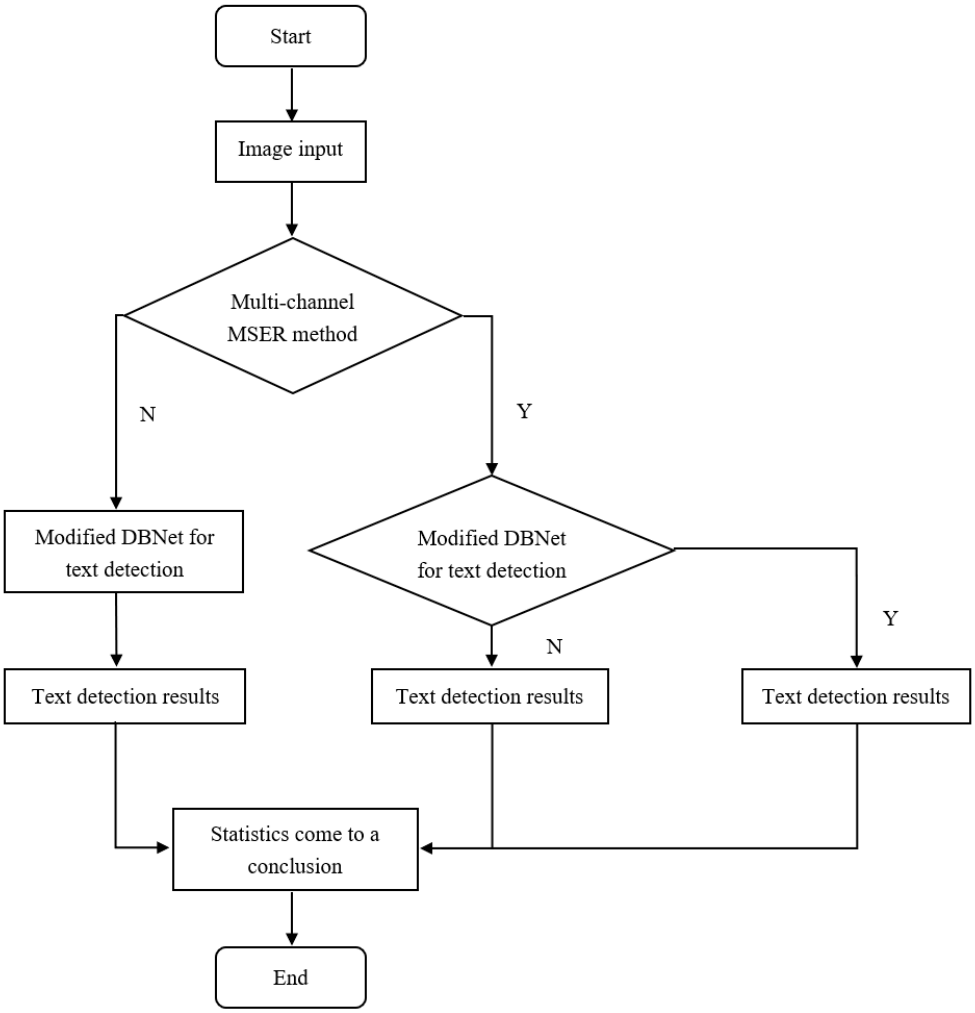


**Figure 7.** Flow chart of the experiment.

The evaluation of natural scene text detection has three main metrics: precision, recall, and F-Score, which are defined as follows:

$$prescision = \frac{TP}{TP + FP} \tag{7}$$

$$recall = \frac{TP}{TP + FN} \tag{8}$$

$$F - score = 2 * \frac{precision * recall}{precision + recall} \tag{9}$$

where *TP* represents the number of samples that are positive and predicted to be positive samples, *FP* represents the number of samples that are negative but predicted to be positive samples, *FN* represents the number of samples that are positive but predicted to be negative samples, and the *F-Score* is a comprehensive metric based on the accuracy and recall, which can better characterize the text detection performance of a network.

### *4.2. Experimental Environment*

The experiments of all the datasets were conducted on a Windows 10 system, developed using Python, and the deep learning framework was PyTorch 1.4.0. The hardware environment included an Intel Corei9 processor (2.80 GHz), 64 GB of memory, and an NVIDIA RTX 2080 SUPER graphics card (8 GB of memory). To improve the training efficiency, all the processed images were adjusted to dimensions of 640 × 640, and the

batch size was set to eight. We adopted the multiple learning rate strategy, that is, the current learning rate was equal to the initial learning rate multiplied by $(1 - \frac{epoch}{\max\_epoch})^{power}$. Combined with the decay characteristics of this learning rate, we set the initial learning rate to 0.007 and the power to 0.9.

### *4.3. Experimental Datasets*

According to the literature survey, so far, no street sign dataset can be used in the natural scene of the experiment. Having a good dataset is the first step for street sign text detection experiments, so in this paper, we created our dataset, NSSS, to validate our proposed method. All of the datasets came from images taken in real environments, including street signs on both sides of daily streets, street signs in schools and shopping malls, street signs in scenic parks, and many other scenes, some data set images are shown in Figure 8. In order to simulate the noise that might be introduced by lighting and weather in natural scenes, some images were randomly adjusted the brightness and contrast. To obtain the true text areas (ground truth) of the images, all the images were manually annotated with text and 2000 images were selected from them for the experiment, of which 1500 were used for the training set and 500 were used for the test set.


**Figure 8.** Partial images of the NSSS dataset.

To further test the effectiveness of our proposed method, the publicly available dataset ICDAR2015 [31] and dataset ICDAR2017-MLT [32] were used for our experiments for evaluation. ICDAR2015 consisted of 1500 images taken in natural scenes and contained a large amount of irregular text. Of these images, 1000 were used as the training set, and 500 were used as the test set. ICDAR2017-MLT is a text dataset composed of multiple languages, including Chinese, English, Korean, Japanese, Italian, French, Indian, Arabic, and German. There are 9000 images in the data set, of which 7200 are used for the training set and 1800 are used for the test set.

### *4.4. Experiment Results and Discussion*

As shown in Table 2, when the HLFE module was added, the high-level semantic information of the network was fully utilized, thereby reducing the phenomenon of missed detection of small-scale text instances and false detection caused by the influence of non-text regions in the model. Therefore, the recall and accuracy were improved to a certain extent. When the FPRE module was added, the low-level semantic information of the network could be better propagated and utilized so that the model could better locate the text location and text boundary, improving the accuracy of the model significantly. When the HLFE and FRPE modules were applied together, the feature extraction effect was further improved. When the detection was performed after processing the image using the multi-channel MSER, the model's recall was improved more significantly because a large number of non-text regions were filtered out, which effectively reduced the influence of non-text regions on the text regions and made the text features more evident, indicating the significance of the MSER method for image pre-processing. The final precision of the model was improved from 91.8% to 92.5%, the recall rate was improved from 82.1% to 86.8%, and the F-Score was improved from 86.7% to 89.5%. As shown in Figure 9, the proposed method in this paper had a better detection effect for regions with inconspicuous text features, and it worked better than the original DBNet model in terms of recall, precision, and F-Score.

**Table 2.** Comparison of results of ablation experiments on the NSSS dataset.

| Methods | Precision (%) | Recall (%) | F-Score (%) |
|---|---|---|---|
| DBNet (baseline) | 91.8 | 82.1 | 86.7 |
| DBNet + HLFE | 91.6 | 83.5 | 87.3 |
| DBNet + FPRE | 92.1 | 83.3 | 87.5 |
| DBNet + HLFE + FPRE | 91.8 | 84.1 | 87.8 |
| DNNet + HLFE + FPRE + MSER (ours) | 92.5 | 86.8 | 89.5 |


**Figure 9.** Comparison of visualization results of the method proposed in this paper and original DBNet model on natural scene street sign (NSSS) dataset: (**a**) is the result of the proposed method in this paper, and (**b**) is the result of the original DBNet model.

The comparison of the effect of the proposed model with the base model on the ICDAR2015 dataset and ICDAR2017-MLT dataset are shown in Figures 10 and 11, indicating that the proposed method had improved in terms of missed detection and false detection compared to the original method. As shown in Tables 3 and 4, the recall and precision of our proposed method on the ICDAR2015 dataset were improved, which indicates that our method is also applicable to other datasets of scenes with text. The proposed method showed good results for different languages, shapes, sizes, scales, and slanted text, and it could cope with text detection in most natural scenes.

**Table 3.** Comparison of results of ablation experiments on the ICDAR2015 dataset.

| Methods | Precision (%) | Recall (%) | F-Score (%) |
|---|---|---|---|
| DBNet (baseline) | 89.5 | 75.9 | 82.1 |
| DBNet + HLFE | 89.3 | 76.7 | 82.5. |
| DBNet + FPRE | 90.1 | 77.1 | 83.1 |
| DBNet + HLFE + FPRE | 88.2 | 78.1 | 83.3 |
| DNNet + HLFE + FPRE + MSER (ours) | 90.8 | 79.0 | 84.5 |

**Table 4.** Comparison of results of ablation experiments on the ICDAR2017-MLT dataset.

| Methods | Precision (%) | Recall (%) | F-Score (%) |
|---|---|---|---|
| DBNet (baseline) | 81.9 | 63.8 | 71.7 |
| DBNet + HLFE | 83.1 | 63.7 | 72.2 |
| DBNet + FPRE | 83.6 | 63.2 | 72.0 |
| DBNet + HLFE + FPRE | 83.2 | 63.9 | 72.3 |
| DNNet + HLFE + FPRE + MSER (ours) | 84.1 | 64.9 | 73.3 |


(a)


(b)

**Figure 10.** Comparison of visualization results between the method proposed in this paper and the original DBNet model on the ICDAR2015 dataset: (**a**) is the result of the proposed method in this paper, and (**b**) is the result of the original DBNet model.


**Figure 11.** Comparison of visualization results between the method proposed in this paper and the original DBNet model on the ICDAR2017-MLT dataset: (**a**) is the result of the proposed method in this paper, and (**b**) is the result of the original DBNet model.

As shown in Figure 12, the method proposed in this paper showed better performances than the Baseline on both the ICDAR2015, ICDAR2017-MLT, and NSSS datasets in that the trained F-score tended to be stable with the increase in the number of training epochs. The yellow and green curves represent the F-score curves of our proposed method and the Baseline of the ICDAR2015 dataset, the purple and cyan curves represent the F-score curves of the method proposed in this paper and the Baseline of the ICDAR2015 dataset, respectively, and the blue and red curves represent the F-score curves of our proposed method and the Baseline on the NSSS dataset, respectively.

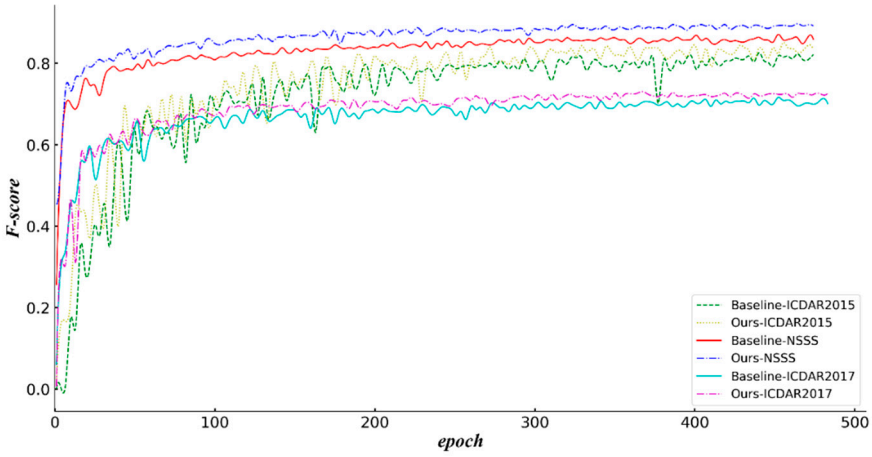


**Figure 12.** F-score curves of Baseline and our method on ICDAR2015, ICDAR2017-MLT, and NSSS datasets.

To verify the effectiveness of the methods proposed in this paper, the following methods were used for a comparative study: the EAST model proposed by Zhou et al. [21], the PixelLink model based on instance segmentation proposed by Deng et al. [15], the rotation-sensitive regression detector (RRD) proposed by Liao et al. [33], the short path network (SPN) proposed by Wang et al. [34], the TextSnake-like detection method (TextSnake) for irregular text proposed by Long et al. [35] and a path aggregation network (PANet) [36] is based on an instance segmentation framework. The objective of these methods was consistent with the goal of the work presented in this paper, which was to detect text in natural scenes.

The comparison of the detection effectiveness of the proposed method with other methods on the ICDAR2015 dataset is shown in Table 5. The precision, recall, and F-Score of the proposed method in this paper were, respectively improved by 7.2%, 5.5%, and 6.3% compared to the EAST. Because the EAST had a small perceptual field and used text candidate frames, it was less effective in detecting long and slanted text. Compared with the segmentation-based PixelLink method, the precision and F-Score were improved by 7.9% and 2.2%, respectively. This was because PixelLink required geometric features as post-processing to detect segmented text and could not effectively handle text regions with complex backgrounds. Compared with the PANet model, the precision and F-Score were also improved by 6.8% and 1.6%, respectively. This was because the PANet did not make full use of the high-level semantic information, resulting in its failure to accurately detect some text regions with large text size differences. The comparison of the detection effectiveness of the proposed method with other methods on the ICDAR2017 dataset is shown in Table 6. Compared with PSENet, the accuracy and F-Score of the proposed method are improved by 7.9% and 2.2%, respectively. Because the PSENet network does not exclude the impact of light and complex backgrounds. Compared with the original DBNet model, the accuracy, recall, and F-Score increased by 2.2%, 1.1%, and 1.6%, respectively. Although our proposed method did not perform as well as some models in terms of recall, it outperformed other methods in terms of precision and F-Score, which fully demonstrated the high precision of our model in classifying textual and non-textual regions of images. Furthermore, it could fully utilize both high-level and low-level semantic information to achieve accurate localization of textual instances.

**Table 5.** Performance comparison of the model proposed in this paper with other models on the ICDAR2015 dataset.

| Methods | Precision (%) | Recall (%) | F-Score (%) |
|---|---|---|---|
| EAST [21] | 83.6 | 73.5 | 78.2 |
| PixelLink [15] | 82.9 | 81.7 | 82.3 |
| RRD [33] | 85.6 | 79.0 | 82.2 |
| SPN [34] | 86.6 | 82.1 | 84.3 |
| TextSnake [35] | 84.9 | 80.4 | 82.6 |
| PANet [36] | 84.0 | 81.9 | 82.9 |
| DBNet [10] | 89.5 | 75.9 | 82.1 |
| Ours | 90.8 | 79.0 | 84.5 |

**Table 6.** Performance comparison of the model proposed in this paper with other models on the ICDAR2017-MLT dataset.

| Methods | Precision (%) | Recall (%) | F-Score (%) |
|---|---|---|---|
| SCUT_DLVlab1 [37] | 80.3 | 54.5 | 65.0 |
| ete_ctc01_multi_scale [37] | 79.8 | 61.2 | 69.3 |
| Corner [37] | 83.8 | 55.6 | 66.8 |
| Zhang et al. [38] | 74.9 | 61.0 | 67.3 |
| PSENet [6] | 77.0 | 68.4 | 72.5 |
| DBNet [10] | 81.9 | 63.8 | 71.7 |
| Ours | 84.1 | 64.9 | 73.3 |

## 5. Conclusions and Feature Works

This paper improved the natural scene street sign text detection method with differentiable binarization networks was proposed and created an NSSS dataset to better support this work. First, to solve the problem that the street sign text is interfered with by complex backgrounds and strong light in natural scenes, the Canny operator was used to enhance the text boundaries and used the multi-channel MSER method to remove a large number of non-text regions, which effectively reduced the interference of non-text regions and strong light on the text detection. In addition, to address the problem that the original DBNet model did not make full use of the high-level and low-level semantic information during the feature extraction network, this paper improved the feature pyramid network of the DBNet model, so that the low-level and high-level semantic information of the network could be more fully utilized, enhancing the ability of the network to detect the text information of street signs. The experimental results showed that the method proposed in this paper achieved significant improvements in the text detection of natural scenes and is quite competitive with existing methods, which proved the effectiveness of the method proposed in this paper.

However, the detection effectiveness of our model decreased when the character spacing of the same text line was large or the image was severely blurred, and the model also could not effectively detect other identifiers in the street signs. In addition, it will be future work to detect text with large character spacings in the same line and other identifiers in street signs that can provide valid information other than text, providing a street sign text detection model with more application value.

**Author Contributions:** The authors' contributions are summarized below. M.L. and Y.L. made substantial contributions to the conception and design. M.L. and Y.L. were involved in drafting the manuscript. M.L., Y.L. and Q.T. acquired data and analysis and conducted the interpretation of the data. A lot of effective suggestions for this manuscript were proposed by C.-L.C. The authors would like to thank the anonymous reviewers and the editors for all the helpful suggestions. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Social Science Fund of China, grant number (20BGL141).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** This study is only based on theoretical basic research. It does not involve human subjects.

**Data Availability Statement:** The data used to support the findings of this study are available from the corresponding author upon request.

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

## References

1. Naosekpam, V.; Sahu, N. Text detection, recognition, and script identification in natural scene images: A Review. *Int. J. Multimedia Inf. Retrieval* **2022**, *11*, 291–314. [CrossRef]
2. Yu, Y.; Jiang, T.; Li, Y.; Guan, H.; Li, D.; Chen, L.; Yu, C.; Gao, L.; Gao, S.; Li, J. SignHRNet: Street-level traffic signs recognition with an attentive semi-anchoring guided high-resolution network. *ISPRS J. Photogramm. Remote Sens.* **2022**, *192*, 142–160. [CrossRef]
3. Guo, J.; Lu, J.; Qu, Y.; Li, C. Traffic-Sign Spotting in the Wild via Deep Features. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; pp. 120–125.
4. Jian, M.; Zhang, W.; Yu, H.; Cui, C.; Nie, X.; Zhang, H.; Yin, Y. Saliency detection based on directional patches extraction and principal local color contrast. *J. Visual Commun. Image Represent.* **2018**, *57*, 1–11. [CrossRef]
5. Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; Bigorda, L.G.; Mestre, S.R.; Mas, J.; Mota, D.F.; Almazan, J.A.; De Las Heras, L.P. ICDAR 2013 Robust Reading Competition. In Proceedings of the 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 1484–1493.
6. Hassan, E. Scene Text Detection Using Attention with Depthwise Separable Convolutions. *Appl. Sci.* **2022**, *12*, 6425. [CrossRef]
7. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.

8. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2015**, arXiv:1511.07122.
9. Matas, J.; Chum, O.; Urban, M.; Pajdla, T. Robust wide-baseline stereo from maximally stable extremal regions. *Image Vision Comput.* **2004**, *22*, 761–767. [CrossRef]
10. 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; p. 34.
11. Naiemi, F.; Ghods, V.; Khalesi, H. A novel pipeline framework for multioriented scene text image detection and recognition. *Expert Syst. Appl.* **2021**, *170*, 114549. [CrossRef]
12. Liu, C.; Yang, C.; Hou, J.; Wu, L.; Zhu, X.; Xiao, L. GCCNet: Grouped channel composition network for scene text detection. *Neurocomputing* **2021**, *454*, 135–151. [CrossRef]
13. Lu, M.; Mou, Y.; Chen, C.L.; Tang, Q. An Efficient Text Detection Model for Street Signs. *Appl. Sci.* **2021**, *11*, 5962. [CrossRef]
14. Wan, Q.; Ji, H.; Shen, L. Self-attention based Text Knowledge Mining for Text Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5983–5992.
15. Deng, D.; Liu, H.; Li, X.; Cai, D. PixelLink: Detecting Scene Text via Instance Segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 6773–6780.
16. Zhu, J.; Wang, G. TransText: Improving scene text detection via transformer. *Digital Signal Processing* **2022**, *130*, 103698. [CrossRef]
17. 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.
18. Hu, Z.; Wu, X.; Yang, J. TCATD: Text Contour Attention for Scene Text Detection. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 1083–1088.
19. Qiao, L.; Tang, S.; Cheng, Z.; Xu, Y.; Niu, Y.; Pu, S.; Wu, F. Text Perceptron: Towards End-to-End Arbitrary-Shaped Text Spotting. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11899–11907.
20. Cai, Y.; Liu, Y.; Shen, C.; Jin, L.; Li, Y.; Ergu, D. Arbitrarily shaped scene text detection with dynamic convolution. *Pattern Recognit.* **2022**, *127*, 108608. [CrossRef]
21. 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.
22. Li, X.; Liu, J.; Zhang, G.; Huang, Y.; Zheng, Y.; Zhang, S. Learning to predict more accurate text instances for scene text detection. *Neurocomputing* **2021**, *449*, 455–463. [CrossRef]
23. Liu, J.; Zhong, Q.; Yuan, Y.; Su, H.; Du, B. SemiText: Scene text detection with semi-supervised learning. *Neurocomputing* **2020**, *407*, 343–353. [CrossRef]
24. He, T.; Huang, W.; Qiao, Y.; Yao, J. Text-Attentional Convolutional Neural Network for Scene Text Detection. *IEEE Trans. Image Proces.* **2016**, *25*, 2529–2541. [CrossRef] [PubMed]
25. Mittal, A.; Shivakumara, P.; Pal, U.; Lu, T.; Blumenstein, M. A new method for detection and prediction of occluded text in natural scene images. *Signal Proces. Image Commun.* **2022**, *100*, 116512. [CrossRef]
26. Xue, M.; Shivakumara, P.; Zhang, C.; Xiao, Y.; Lu, T.; Pal, U.; Lopresti, D.; Yang, Z. Arbitrarily-Oriented Text Detection in Low Light Natural Scene Images. *IEEE Trans. Multimedia* **2020**, *23*, 2706–2720. [CrossRef]
27. Ding, L.; Goshtasby, A. On the Canny edge detector. *Pattern Recognit.* **2001**, *34*, 721–725. [CrossRef]
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
29. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable ConvNets V2: More Deformable, Better Results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 5–20 June 2019; pp. 9308–9316.
30. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
31. Karatzas, D.; Gomez, B.; Nicolaou, A. ICDAR 2015 competition on Robust Reading. In Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 1156–1160.
32. Iwamura, M.; Morimoto, N.; Tainaka, K.; Bazazian, D.; Gomez, L.; Karatzas, D. ICDAR2017 Robust Reading Challenge on Omnidirectional Video. In Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; pp. 1448–1453.
33. Liao, M.; Zhu, Z.; Shi, Z.B. Rotation-Sensitive Regression for Oriented Scene Text Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5909–5918.
34. Cai, Y.; Wang, W.; Ren, H.; Lu, K. SPN: Short path network for scene text detection. *Neural Comput. Appl.* **2019**, *32*, 6075–6087. [CrossRef]
35. Long, S.; Ruan, J.; Zhang, W.; He, X.; Wu, W.; Yao, C. TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes. In Proceedings of the European Conference on Computer Vision (ECCV), Cham, Switzerland, 9 October 2018; pp. 20–36.
36. 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 International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8440–8449.

37. 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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7553–7563.
38. Zhang, S.X.; Zhu, X.; Hou, J.B.; Liu, C.; Yang, C.; Wang, H.; Yin, X.C. Deep Relational Reasoning Graph Network for Arbitrary Shape Text Detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9696–9705.