Text Spotting of Electrical Diagram Based on Improved PP-OCRv3

Yuqian Zhao[1], Dongdong Zhang[1](✉), and Chengyu Sun[2]

[1] Department of Computer Science and Technology, Tongji University,
Shanghai, China
{2233047,ddzhang}@tongji.edu.cn
[2] Shanghai Key Laboratory of Urban Renewal and Spatial Optimization Technology,
Tongji University, Shanghai, China
cy.sun@tongji.edu.cn

Abstract. The text detection and recognition plays an important role in automatic management of electrical diagrams. However, the images of electrical diagrams often have high resolution, and the format of the text in them is also unique and densely distributed. These factors make the general-purpose text spotting models unable to detect and recognize the text effectively. In this paper, we propose a text spotting model based on improved PP-OCRv3 to achieve better performance on text spotting of electrical diagrams. Firstly, a region re-segmentation module based on pixel line clustering is designed to correct detection errors on irregularly shaped text containing vertical and horizontal characters. Secondly, an improved BiFPN module with channel attention and depthwise separable convolution is introduced during text feature extracting to improve the robustness of input images with different scales. Finally, a character re-identification module based on region extension and cutting is added during the text recognition to reduce the adverse effects of simple and dense character on the model. The experimental results show that our model has better performance than the state-of-the-art (SOTA) methods on the electrical diagrams data sets.

Keywords: Electrical diagrams · Optical character recognition · PP-OCRv3 · BiFPN

1 Introduction

It is urgent to automate the management and search of electrical diagrams. As the scale of electrical diagrams expands, traditional manual identification is not only inefficient but also prone to errors. The information in the electrical diagrams is complex. In order to realize the automation process, it is important to detect the text position and recognize the text content in the electrical diagrams.

In recent years, deep learning based text detection and recognition methods have been proposed. These methods hand over the classification features of

B. Luo et al. (Eds.): ICONIP 2023, CCIS 1968, pp. 125–136, 2024.
https://doi.org/10.1007/978-981-99-8181-6_10

electrical diagrams automatically to the networks, thereby improving the adaptability to complex scenarios of electrical diagrams. Laura Jamieson et al. [1] used EAST and LSTM to realize the detection and recognition of text in engineering diagrams. Li et al. [2] used YOLOv5 and PP-OCRv3 [3] to improve the accuracy on electrical cabinet wiring.

However, the forms of text labeling in the electrical diagrams are often diverse, with different shapes, directions, and sizes. The characters involved not only numbers and letters, but also Chinese characters. The accuracy is often poor when directly using the general-purpose OCR model to spotting the text in the electrical diagrams. Figure 1 shows the text detection results of electrical diagrams on PP-OCRv3, where four types of text are prone to errors and omissions in detection and recognition.

To address these limitations and achieve better position detection and content recognition of text in electrical diagrams, we propose a text spotting model based on PP-OCRv3. To improve the detection performance for irregularly shaped text containing both vertical and horizontal characters, a region re-segmentation module based on pixel line clustering is designed. During text feature extraction, we propose an improved BiFPN module with channel attention and depthwise separable convolution to improve the problem of mis-identification caused by text images with different scales. To correct the error of texts with single Chinese character and multi-line numbers, a character re-identification module based on region extension and cutting is introduced, which could improve performance of the model in electrical diagrams. To summarize, our main contributions are as follows:

- We design a region re-segmentation module based on pixel line clustering to solve the text position detection problem caused by irregularly shaped text containing both vertical and horizontal characters.
- We propose an improved BiFPN module with channel attention and depthwise separable convolution for text feature extraction to improve the accuracy of texts with different scales.
- We introduce a charater re-identification module based on region extension and cutting to correct the wrongly text recognition caused by texts with simple Chinese charaters and multi-line numbers.
- Extensive experiments show that our method has better performance than the SOTA methods on the electrical diagrams data sets.

2 Related Work

In the field of OCR technology, some models are proposed to detect the text position. With the introduction of image segmentation, most of the text detection models can achieve the task of detecting arbitrarily shaped text, but the detection capability for dense texts is poor. As a result, PSENet [4] and PAN [5] were proposed to improve the performance of text detection on compact text detection. DBNet [6] introduced a boundary threshold selection strategy to achieve accurate detection results.

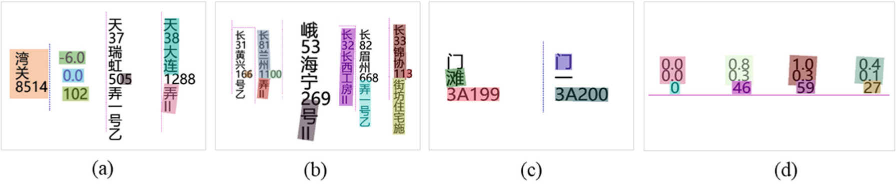


Fig. 1. Some text instances that are hard to detect and recognize using PP-OCRv3.(a) Irregularly shaped texts that contain both horizontal and vertical characters. (b) Texts with various sizes and shapes. (c) Texts that contain simple Chinese characters, which is easily omitted. (d) Texts that contain dense multi-line numbers.

Other models are proposed to recognize the text content. With the emergence of sequence-to-sequence [7], text recognition models for the extraction and analysis of text sequence features begin to emerge, the most representative of which is CRNN [8]. Sheng et al. [9] used transformer to encode and decode input images, and only use several simple convolutional layers for high-level feature extraction, which verified the effectiveness of transformer in text recognition. Du et al. [10] proposed SVTR, which is based on global and local self-attention mechanisms to obtain large-scale correlation between characters and small-scale correlation between stroke structures, respectively.

In recent years, one-stage text detection and recognition has become the main focus of research in OCR. Most of the one-stage models, like MANGO [11], Mask TextSpotter v3 [12] and PGNet [13], combine the network structure into one overall architecture, which can effectively reduce the inconsistency of detection. However, most of these one-stage models are focused on character-level recognition, which is not applicable to more complex scenes.

PP-OCRv3 is a representative one of the two-stage text spotting models. During text position detection, DBNet is introduced to detect the text regions. And then the text proposals are obtained. During the text feature extraction, cropped text images according to the text proposals are fed into MobileNetv1 to extract features. The feature maps Fs of each text proposals is obtained. During text recognition, a recognition head based on SVTR is proposed to obtain the text recognition result.

However, as a general text spotting model, PP-OCRv3 performs poorly for multi-scale electrical diagrams with irregularly shaped and compact distributed texts. To enhance the performance on text spotting of electrical diagrams, we propose a text spotting model of electrical diagrams based on improved PP-OCRv3.

3 Methodology

In order to solve the problem of text spotting of electrical diagrams, we propose text spotting model of electrical diagrams based on improved PP-OCRv3, which is shown in Fig. 2. Our model includes text position detection, text feature extraction and text recognition. The blue part is the module in the original PP-OCRv3, and the green part is the module we proposed. In order to enhance the detection accuracy for irregularly shaped text, we design a region re-segmentation module based on pixel row clustering in text position detection. In order to improve the robustness for text images of different scales, we introduce an improved BiFPN module in text feature extraction. To correct the detection and recognition errors of simple Chinese characters and multi-line numbers, we add a character re-identification module based on region extension and cutting in text recognition.

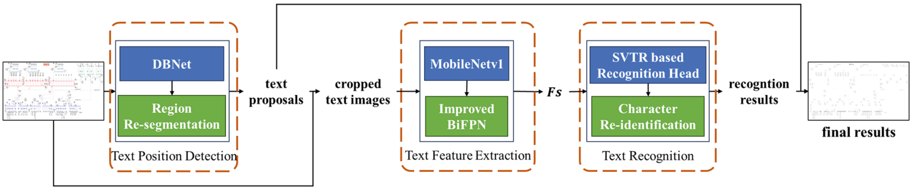


Fig. 2. The overall structure of text spotting model of electrical diagrams.

3.1 Text Position Detection Based on DBNet and Region Re-segmentation

The overall framework of the text position detection is shown in Fig. 3. We follow PP-OCRv3 to use DBNet as the detection network, as shown in the green boundary area. Since the electrical diagrams contain irregularly shaped texts that contain both horizontal and vertical characters, the segmentation results of existing methods are not good. Therefore, we design a region re-segmentation module based on pixel line clustering after DBNet. As shown in Fig. 3, the gray part is the wrong segmentation result of the original PP-OCRv3, and the yellow part is the region re-segmentation module we designed.

The specific steps of the region re-segmentation module are as follows.

Step 1: Unusual regions filtering. Calculate the IoUs of each kernel contour and kernel box. The kernel contour is tight and non-rectangular to express the outline of the text segmentation area, while the kernel box is loose and rectangular to express the bounding rectangle of the text segmentation area. $IoU < \theta$ often means that the text kernel has an irregular shape and is regarded as an "unusual region", where θ is a hyper-parameter.

Step 2: Pixel line clustering. Record the width of each pixel line of the unusual contour region. After that, perform line-based clustering on the contour area.

Step 3: Threshold filtering. Set δ as a threshold of the area. Filter the class cluster if the area of it is less than δ. In this way, clusters that are too small to contain text are discarded, while the proper clusters then become the fine-tuned kernels.

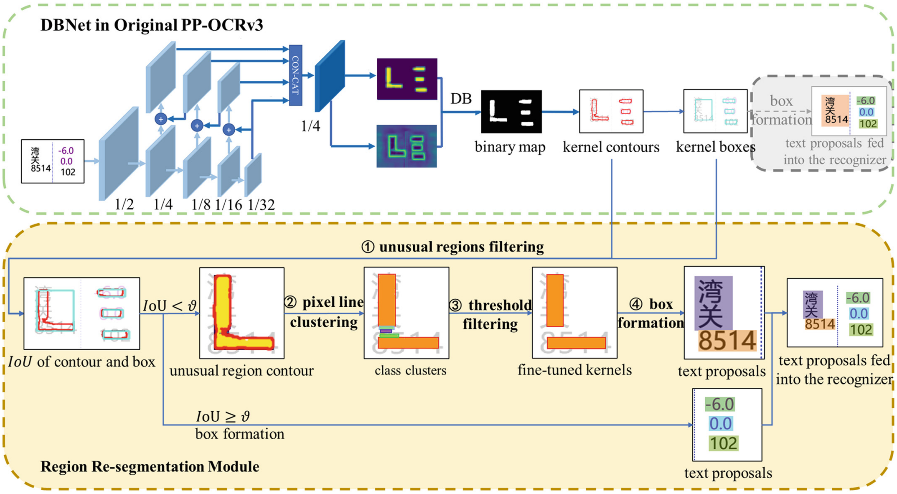


Fig. 3. The structure of text poition detection.

Step 4: Box formation. Generate text proposals according to fine-tuned kernels.

3.2 Text Feature Extraction Based on MobileNetv1 and Improved BiFPN

Our proposed text feature extraction network is designed on the basis of MobileNetv1 of PP-OCRv3. An improved BiFPN module with channel attention and depthwise separable convolution is introduced after MobileNetv1 to extract multi-scale features of texts with different shapes in electrical diagrams. Specifically, a channel attention block is added at the end of BiFPN to enhance information in high-level feature maps and improve the global feature fitting ability. In order to overcome the increase of parameters, depthwise separable convolutions is used to replace all convolutions in the network. Hardswish is uesd as the activation function to further improve the accuracy.

The overall network structure is shown in Fig. 4. The blue area is the structure of the original feature extraction network MobileNetv1 in PP-OCRv3. The feature maps F_1, F_2, F_3, and F_4 are the outputs of different layers of MobileNetv1, and the scales are $64 \times 24 \times 160$, $128 \times 12 \times 160$, $256 \times 6 \times 160$, $512 \times 3 \times 160$. The above feature maps are normalized to 256 channels, which is shown as follows:

$$F_i^{in} = HS(Conv_{1\times1}(F_i)) \tag{1}$$

where i is from 1 to 4, and HS is activation function hardswish. Up-sampling is then performed using nearest neighbor interpolation, and the up-sampled results are weighted and summed from top to bottom. The top-down fusion process is as follows:

$$F_i^{mid} = DSConv(w_i^1 \cdot F_i^{in} + w_i^2 \cdot Upsample(F_{i+1}^{in})) \tag{2}$$

where i is from 1 to 3, $DSConv$ is the depthwise separable convolution, and w_i^1 and w_i^2 are the trainable weights of the components of layer i in the top-down fusion process. Maximum pooling is then used for down-sampling, and the down-sampled results are weighted and summed from bottom to top. The bottom-up fusion process is as follows:

$$F_i^{out} = DSConv({w'}_i^1 \cdot F_i^{mid} + {w'}_i^2 \cdot Downsample(F_{i-1}^{mid}) + {w'}_i^3 \cdot F_i^{in}) \tag{3}$$

where i is from 2 to 4, and ${w'}_i^1$, ${w'}_i^2$ and ${w'}_i^3$ are the trainable weights of the components of layer i in the bottom-up fusion process. After that, the aggregated top-level feature map F_4^{out} is down-sampled by a $DSConv$ to reduce the scale to 3×80 and change the number of channels from 256 to 512.

Finally, the feature map is passed through a channel attention module. The structure of the channel attention module is shown in Fig. 5. First, adaptive global average pooling is performed. Then feed it into two fully connected layers with a convolution kernel size of 1×1 to achieve non-linearity, and a weight sequence of $512 \times 1 \times 1$ is obtained. This sequence is used to represent the contribution of each channel of the input feature map. Finally, the sequence of weights is multiplied with the input feature map. The obtained result is a feature map improved by the channel attention mechanism. The specific computational procedure is as follows:

$$F_{out} = F_{in} \cdot (sigmoid(Conv_{1\times1}(ReLU(Conv_{1\times1}(AdaptiveAvgPool(F_{in})))))) \tag{4}$$

After down-sampling, we get a sequence feature map F with the scale of $512 \times 1 \times 40$.

3.3 Text Recognition Based on SVTR and Character Re-identification

The structure of text recognition is shown in Fig. 6. We continue to use the recognition head based on SVTR in PP-OCRv3, as shown in the green area. In order to solve the omission of simple Chinese characters and the wrong recognition of multi-line numbers caused by the original model, we design a character re-identification module based on region extension and cutting behind the recognition head, as shown in the blue region in Fig. 6.

The specific steps of the character re-identification module are as follows.

Step 1: Unusual results filtering. Filter the results output by the recognition header. If the recognition result contains only one Chinese character, then go to step 2. If the recognition result only contains numbers and punctuation, and the text proposal corresponding to the recognition result is square, then go to step 3. We consider the rest of the recognition results to be correct.

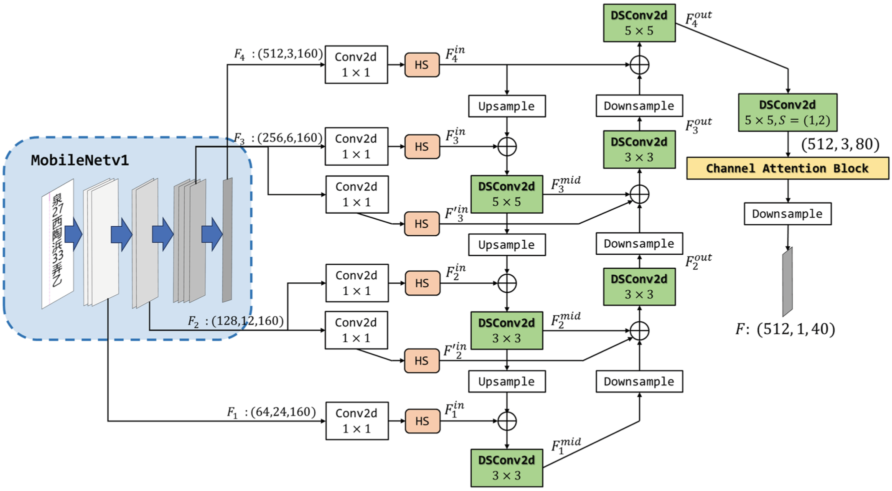


Fig. 4. The framework of text feature extraction network.

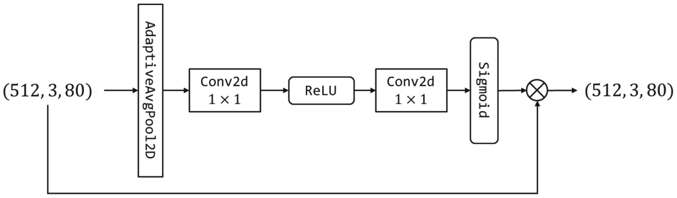


Fig. 5. Channel Attention Block.

Step 2: Vertical extension. We start from the position of the recognized single Chinese character, and extend vertically to the direction of “with text pixels” but “without text proposals” according to the original image and binary map.

Step 3: Horizontal cutting. We cut the text proposals corresponding to the wrong recognition results into two parts horizontally.

Step 4: Re-identification. We feed the corrected text regions in step 2 and step 3 to the text feature extraction and recognition head for re-identification. Finally, we combine the text re-identification results with the previous correct results to get the final recognition result, as shown in the yellow boundary area in Fig. 6.

4 Experiments

4.1 Datasets and Settings

Datasets. When we train the model, we use ICDAR2019-LSVT, ICDAR2017-RCTW and Chinese-Dataset with 127,727 text images as the training set and

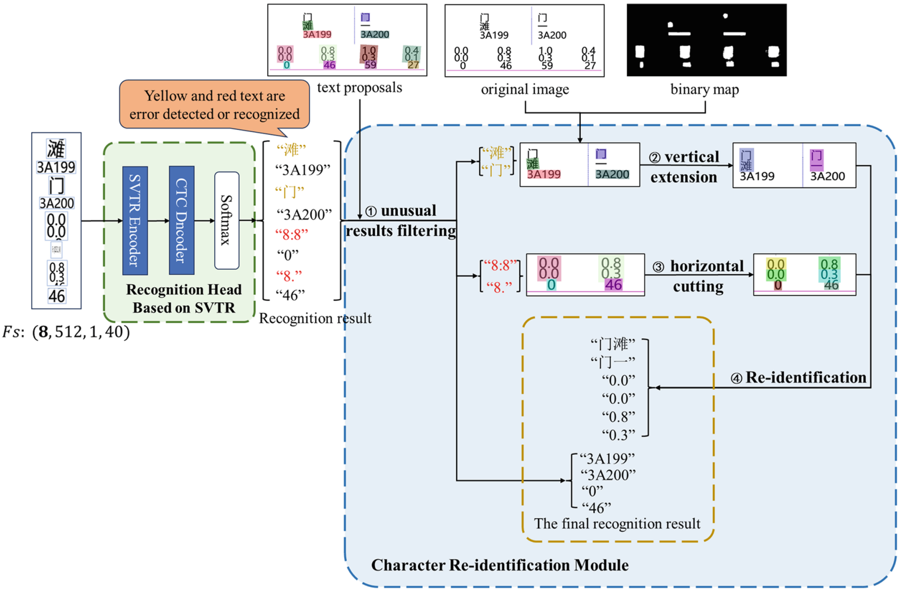


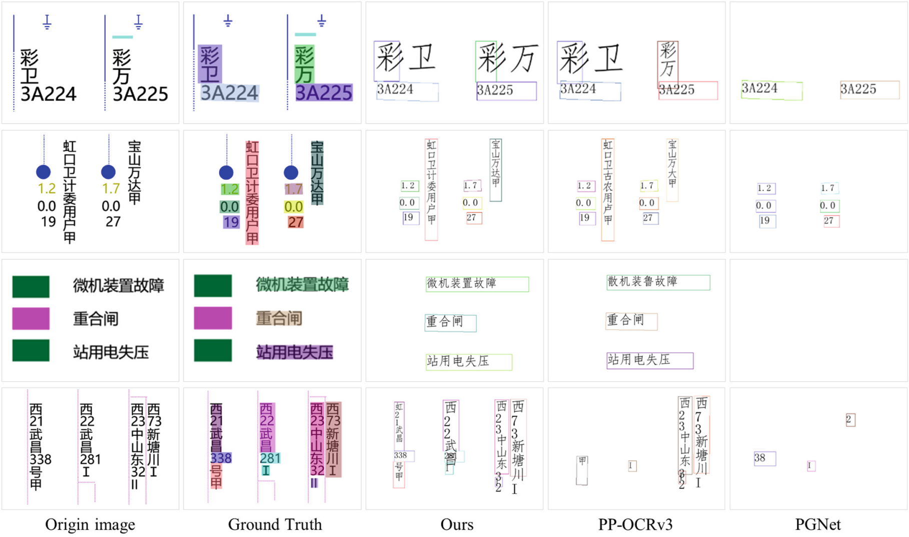
Fig. 6. The framework of text recognition.

ICDAR2015 with 5,230 text images as the validation set for general scene text recognition. To improve the recognition of text in electrical diagrams, we add 23,377 text images of electrical diagrams from State Grid Shanghai Municipal Electric Power Company in the training and validation process.

Settings. During training, we use Adam optimizer with a learning rate of 0.001 and a batch size of 64. We reshape each text instance to 48 × 320 pixels. The threshold θ and δ in text position detection are set to 0.6 and 3.0, respectively. All experiments are performed in PaddlePaddle framework on RTX 3090 GPU. The main results of the experiments including comparison with prior methods and ablation studies, are presented in the following sections.

4.2 Comparison with Prior Methods

Qualitative Evaluation. We show the performance of a one-stage model PGNet [13], PP-OCRv3 and our model in electrical diagrams in Fig. 7. As an efficient one-stage end-to-end text spotting model, the character prediction module of PGNet does not have the sequence encoding ability or the global information acquisition capability. Therefore, PGNet only performs well for recognition of numbers and letters, and cannot achieve recognition of Chinese (as shown in rows 1, 2 and 4 of Fig. 7, it can only recognize numbers and letters effectively). PP-OCRv3 can effectively encode sequence information, and thus has the ability to recognize long Chinese text. However, it is not adaptable to scale changes, resulting in a poor ability to capture texture features in large scale images (as shown in rows 2 and 3 of Fig. 7, with poor performance for long text). In addition, the detection accuracy of PP-OCRv3 for irregular compact text is low due to the lack of targeted fine-tuning operations (as shown in row 4 of Fig. 7, irregularly shaped compact text cannot be detected). Comparatively speaking, our model can be more adapted to the characteristics of electrical diagrams, and has better recognition performance for multi-scale text images.



Origin image Ground Truth Ours PP-OCRv3 PGNet

Fig. 7. Examples of results for comparison.

Table 1. Quantitative comparison over different methods.

Method	Detection			Character Accuracy	End-to-End		
	P	R	F1		P	R	F1
PGNet	72.93	70.86	71.88	62.46	55.75	51.39	53.48
EAST+CRNN [8]	83.30	89.52	86.30	68.80	62.12	67.26	64.59
PSENet [4]+NRTR [9]	86.86	86.12	86.98	70.01	68.96	66.80	67.86
ABCNet v2 [14]	89.42	87.24	88.32	72.46	70.74	67.38	69.02
SwinTextSpotter [15]	90.18	88.02	89.05	76.55	73.87	69.82	71.79
PP-OCRv3	86.23	93.43	89.68	80.84	71.89	77.90	74.77
Ours	**90.38**	**95.56**	**92.90**	**85.39**	**78.27**	**82.76**	**80.45**

Quantitative Evaluation. To further compare the accuracy of our model with the SOTA approach, we perform quantitative evaluation on electrical diagram test set, as shown in Table 1.

The column *Detection* shows the accuracy of text position detection. We use P, R and $F1$ as the evaluation indexes of *Detection*. The column $CharacterAccuracy$ indicates the percentage of text images that are correctly recognized during text feature extraction and text recognition. The column $End-to-End$ shows the text detection and recognition accuracy of the whole system.

As a one-stage model, PGNet's character recognition module does not have the capability of sequence encoding and global information extraction, so it can only recognize letters and numbers, but cannot recognize Chinese sequences. The next five models are representative two-stage models. The detector and the recognizer use different models or adopt different optimization strategies. Among them, PP-OCRv3 shows the best performance due to the application of differentiable binarization and global/local mixing block. Therefore, we choose PP-OCRv3 as the baseline. According to above comparison, our model improves by 5.68% in terms of $F1$ over PP-OCRv3. The results show that our model achieves the best performance.

4.3 Ablation Study

Analysis on Region Re-segmentation Module. We compare the test results with and without the region re-segmentation module to verify the impact of it. As shown in Fig. 8, without the re-segmentation module, the model performs poorly for the detection and recognition of irregular text and text containing both vertical and horizontal characters.

Analysis on Improved BiFPN Module. We compare the test results with and without the improved BiFPN module to verify the impact of it, as shown in Fig. 8. The introduction of improved BiFPN enables the recognition branch to focus on the contour information of the characters as well as the texture information, which is more important for text recognition.

Analysis on Character Re-identification Module. We compare the test results with and without the character re-identification module to verify the impact of it. Without the re-identification module, the model has poor location ability for characters with simple structures and numbers with compact positions, which can lead to some missed or incorrect detection, as shown in Fig. 8.

The ablation study results of each module are shown in Table 2. It can be seen that our proposed strategy enables the model to effectively adapt to the actual recognition needs of text in electrical diagrams.

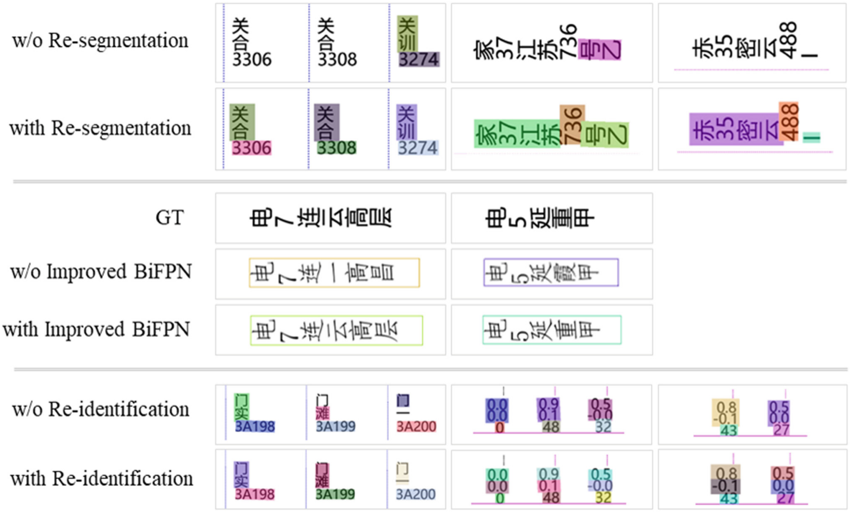


Fig. 8. Ablation study.

Table 2. Ablation study.

Improved BiFPN	Re-segmentation	Re-identification	Detection			Character Accuracy	End-to-End		
			P	R	F1		P	R	F1
✗	✗	✗	86.23	93.43	89.68	80.84	71.89	77.90	74.77
	✗	✗	86.56	93.54	89.91	82.20	74.20	80.17	77.07
		✗	86.56	94.50	90.36	83.09	75.61	80.61	78.03
✓	✓	✓	**90.38**	**95.56**	**92.90**	**85.39**	**78.27**	**82.76**	**80.45**

5 Conclusion and Future Work

In this paper, we propose a text spotting model of electrical diagrams based on improved PP-OCRv3 to enhance the accuracy of text detection and recognition in electrical diagrams. Firstly, we design a region re-segmentation module based on pixel line clustering to help the model better adapt to the irregularly shaped text containing both vertical and horizontal characters. Secondly, we introduce an improved BiFPN module with channel attention and depthwise separable convolution to achieve high-performance recognition of different scales of texts in electrical diagrams. Finally, to correct the mis-identification of simple Chinese characters and compact multi-line numbers, a character re-identification module based on region extension and cutting is used to help the model better adapt to the complex scenarios. Comprehensive experiments show that our model can significantly improve the performance of text spotting and can be effectively applied to the specific text spotting scene of electrical diagrams. For future work, semantic knowledge in electrical diagrams is further fused to improve the overall performance of the model.

References

1. Jamieson, L., Moreno-Garcia, C.F., Elyan, E.: Deep learning for text detection and recognition in complex engineering diagrams. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE (2020)
2. Shanbin, L., Haoyu, W., Junhao, Z.: Electrical cabinet wiring detection method based on improved yolov5 and pp-ocrv3. In: 2022 China Automation Congress (CAC), pp. 6503–6508. IEEE (2022)
3. Li, C., et al.: Pp-ocrv3: more attempts for the improvement of ultra lightweight OCR system. arXiv preprint arXiv:2206.03001 (2022)
4. Wang, W., et al.: Shape robust text detection with progressive scale expansion network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9336–9345 (2019)
5. Wang, W., et al.: Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8440–8449 (2019)
6. 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, vol. 34, pp. 11474–11481 (2020)
7. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. Adv. Neural Inf. Process. Syst. **27** (2014)
8. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(11), 2298–2304 (2016)
9. Sheng, F., Chen, Z., Xu, B.: NRTR: a no-recurrence sequence-to-sequence model for scene text recognition. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 781–786. IEEE (2019)
10. Du, Y., et al.: SVTR: scene text recognition with a single visual model. arXiv preprint arXiv:2205.00159 (2022)
11. Qiao, L., et al.: Mango: a mask attention guided one-stage scene text spotter. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2467–2476 (2021)
12. Liao, M., Pang, G., Huang, J., Hassner, T., Bai, X.: Mask TextSpotter v3: segmentation proposal network for robust scene text spotting. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12356, pp. 706–722. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58621-8_41
13. Wang, P., et al.: Pgnet: real-time arbitrarily-shaped text spotting with point gathering network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2782–2790 (2021)
14. Liu, Y., et al.: Abcnet v2: adaptive Bezier-curve network for real-time end-to-end text spotting. IEEE Trans. Pattern Anal. Mach. Intell. **44**(11), 8048–8064 (2021)
15. Huang, M., et al.: Swintextspotter: scene text spotting via better synergy between text detection and text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4593–4603 (2022)