Check for updates

PCB CT Image Element Segmentation Model Optimizing the Semantic Perception of Connectivity Relationship

Chen Chen, Kai Qiao, Jie Yang, Jian Chen and Bin Yan*

School of Information Systems Engineering, PLA Strategy Support Force Information Engineering University, Zhengzhou, 450001, China

*Corresponding Author: Bin Yan. Email: ybspace@hotmail.com

ABSTRACT

Computed Tomography (CT) is a commonly used technology in Printed Circuit Boards (PCB) non-destructive testing, and element segmentation of CT images is a key subsequent step. With the development of deep learning, researchers began to exploit the "pre-training and fine-tuning" training process for multi-element segmentation, reducing the time spent on manual annotation. However, the existing element segmentation model only focuses on the overall accuracy at the pixel level, ignoring whether the element connectivity relationship can be correctly identified. To this end, this paper proposes a PCB CT image element segmentation model optimizing the semantic perception of connectivity relationship (OSPC-seg). The overall training process adopts a "pre-training and fine-tuning" training process. A loss function that optimizes the semantic perception of circuit connectivity relationship (OSPC Loss) is designed from the aspect of alleviating the class imbalance problem and improving the correct connectivity rate. Also, the correct connectivity rate index (CCR) is proposed to evaluate the model's connectivity relationship recognition capabilities. Experiments show that mIoU and CCR of OSPC-seg on our datasets are 90.1% and 97.0%, improved by 1.5% and 1.6% respectively compared with the baseline model. From visualization results, it can be seen that the segmentation performance of connection positions is significantly improved, which also demonstrates the effectiveness of OSPC-seg.

KEYWORDS

Semantic segmentation; PCB non-destructive testing; mask image modeling; connectivity relationship

1 Introduction

As electronic devices become increasingly complex, Printed Circuit Boards (PCBs), which play a role in mechanical support and electrical connection, are becoming increasingly important for the normal operation of equipment. In the daily inspection and failure analysis of PCBs, the use of Computed Tomography (CT) technology to perform non-destructive testing is a practical method. The main elements in PCBs include pads, wires, and vias, which realize the main electrical connection functions. Therefore, the accurate extraction and segmentation of these elements is crucial, which may affect the accuracy of the subsequent steps of non-destructive testing. Element segmentation belongs to semantic segmentation in computer vision. Before the widespread application of deep learning, element segmentation generally used traditional image segmentation algorithms to perform

single-element segmentation based on the characteristics of a certain element. With the development of deep learning, researchers have begun to use supervised semantic segmentation algorithms for single-element segmentation or component segmentation. However, due to the difficulty in obtaining element-level labeled data, there are few segmentation methods for multiple elements. In order to solve the above problems, self-supervised pre-training has been introduced into this field [1], related methods such as the Contrastive Dual-Masked Autoencoder Pretraining Model for PCB CT Image Element Segmentation (CD-MAE) [1] and efficient pretraining model based on multi-scale local visual field feature reconstruction for PCB CT image element segmentation (EMLR-seg) [2] have emerged. These methods use unlabeled data in the pre-training stage and only a small amount of labeled data in the fine-tuning stage. It generally uses a transformer network with a large number of parameters, greatly improving the robustness of the model [1]. However, in non-destructive testing of circuit boards, it is necessary not only to accurately identify the size and position of elements but also to detect the connectivity between elements, because it will directly affect the subsequent analysis of electrical connections [3] (Fig. 1 shows some disconnections and misconnections in EMLR-seg's segmentation results). But as shown in Fig. 2 the background accounts for a large proportion of PCB images, and the number of pixels of different elements in a dataset can also vary greatly. The class imbalance problem during segmentation will affect the identification of connectivity. Also, the commonly used cross entropy loss only focuses on the overall accuracy of segmentation, completely ignoring whether the model correctly identifies the connectivity relationship.
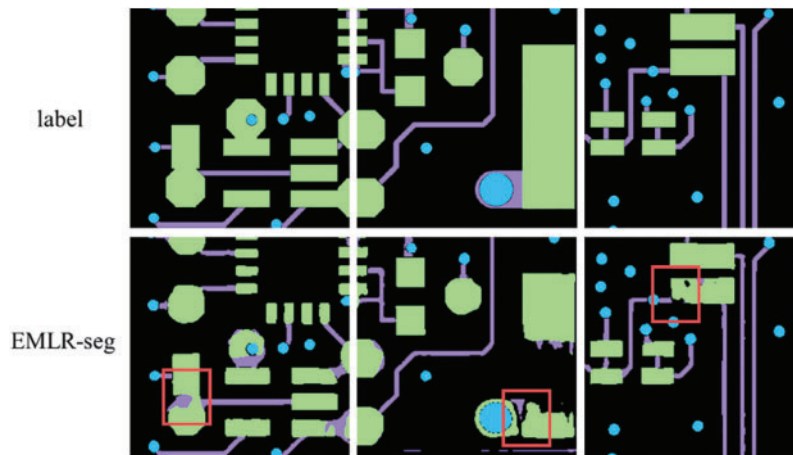

Figure 1: Segmentation results of the baseline model EMLR-seg. There are disconnections and misconnections in the connected locations
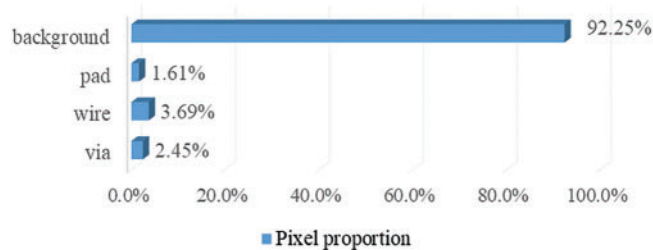

Figure 2: The pixel proportion of each class in our dataset. The background has the largest pixel proportion, and the pad has the smallest pixel proportion, which indicates a serious class imbalance

In response to the above issues, we propose a PCB CT image element segmentation model optimizing the semantic perception of connectivity relationship (OSPC-seg). This method adopts a "pre-training and fine-tuning" training process, mainly improving the semantic perception ability of circuit connectivity relations from the aspect of the loss function. In pre-training, we use EMLR-seg, a pre-training method with good performance in this field. In fine-tuning, we remove the pre-trained decoder and replace the decoder with the UperNet [4] segmentation head. The loss in the fine-tuning stage is divided into two sub-loss functions. In order to alleviate class imbalance, we design a Dynamic IoU-Weighted Focal Loss (DIW Focal Loss) to dynamically control the weights of each class of Focal Loss from the aspect of the evaluation index. In order to directly improve the correct connectivity rate in the segmentation results, a Connectivity Rate sub-loss (CR Loss) is designed. The linear addition of the two yields the loss for optimizing the semantic perception of circuit connectivity (OSPC Loss), and based on this loss, the correct connectivity rate (CCR) is proposed as the evaluation index for connectivity recognition ability. Experiments show that mIoU and CCR of OSPC-seg are 90.1% and 97.0%, improved by 1.5% and 1.6% respectively compared with the baseline model. From visualization results, it can be seen that the segmentation performance of connection positions is significantly improved, which also demonstrates the effectiveness of OSPC-seg.

2 Related Work

2.1 PCB CT Image Element Segmentation

Before the introduction of deep learning, PCB CT image element segmentation could only target a single element class. For example, Atherton et al. [5] proposed the use of the Hough transform to detect vias and Yin et al. [6] proposed the use of the graph cut method to detect vias. These methods are limited by the difficulty of algorithm parameter design and the number of parameters. With the development of deep learning, the use of supervised models trained with labeled data has greatly improved the segmentation accuracy of a single element class or components [3,7,8]. However, the segmentation performance of multiple elements is still limited by the difficulty of obtaining labels and the poor robustness of models. After the "pre-training and fine-tuning" training process is widely used in computer vision, multi-class segmentation models begin to appear, such as CD-MAE proposed by Song et al. [1], which improves feature extraction capabilities by combining contrastive learning, and EMLR-seg proposed by Chen et al. [2] to improve training efficiency by extracting local windows. The evaluation index of these methods is mIoU, which represents the overall segmentation accuracy. However, in non-destructive testing, the recognition of the connectivity relationship between elements is very important, directly affecting the subsequent analysis of electrical connections. Therefore, it is necessary to propose a new evaluation index for connectivity rate. In order to improve the model's perception of connectivity relationship semantics, it is also necessary to design a new loss function to guide model training.

2.2 Masked Image Modeling (MIM)

MIM [9] is a self-supervised method that can be used as a pre-training task. This method applies a certain percentage of the mask to images, then uses an encoder to learn the visible patch features, and a decoder to reconstruct the original image, continuously optimizing the model's ability to learn general features. The reconstruction target has many forms, such as pixels [10,11], discretized tokens [12,13], gradient histograms [14], features generated by other networks [15–18] and features extracted by a teacher network [19]. Currently, the method with better performance and efficiency in the natural image field is the last one. It generally includes a teacher network and a student network. The teacher

network generates the reconstruction target, and the student network performs the reconstruction task. Compared with pixels, it reduces the influence of low-level semantics such as color and texture and is more suitable for PCB CT images, which have metal artefacts, uneven grayscale, translucent and other noise. Also, the synchronously trained teacher network is the only method that does not require additional parameters and training stages to generate reconstruction targets [19].

2.3 Loss Function to Alleviate the Class Imbalance Problem in Semantic Segmentation

There are two main types of loss functions for solving the class imbalance problem. One type is based on the Cross-Entropy loss function (CE Loss) [20,21]. It is a commonly used loss function for semantic segmentation, which is a measure of the difference between two distributions. Focal Loss [22] explains class imbalance from the aspect of difficult samples and adds hyperparameters to adjust the weights of difficult samples in CE Loss.

The other is the IoU-type loss function. Since the final evaluation index of segmentation is IoU, it is possible to optimize the model training directly through IoU Loss [23]. Lovász Loss [24] approximates IoU by constructing a differentiable loss function. The IoU-type loss function has the problem of unstable training, and the class weights of multi-class Focal Loss generally use fixed values, which is not effective in multiple rounds of backpropagation. Therefore, trying to find a method to dynamically change the class weights, and designing the loss function combined with the advantages of IoU Loss are necessary.

3 Method

The overall process of OSPC-seg adopts the "pre-training and fine-tuning" training process. In pre-training, the EMLR-seg pre-training model is used. The fine-tuning model splices part of the UperNet structure as the segmentation head based on the pre-trained encoder. In order to improve the segmentation model's semantic perception of connectivity, the OSPC Loss function is proposed, and CCR is proposed as an evaluation index of connectivity semantic perception ability.

3.1 Pre-Training

The pre-training network includes a teacher network and a student network. As shown in Fig. 3, the student network contains a 24-layer transformer encoder and an 8-layer decoder, and the teacher network only contains a 24-layer transformer encoder. The default size of the input image is 224×224, which is divided into equal and non-overlapping 16×16 image patches, so a complete image has a total of 14×14 image patches. The image patches are embedded through a linear projection, and then several windows of 5×5, 7×7 and 9×9 image patches are randomly selected and input into the teacher network to generate the reconstruction target vectors. At the same time, a certain proportion of the same image is masked, and the left visible patches are input into the student network. The encoder learns the visible patch features. Then encoded visible patches and masked patches are arranged in the original order, and the same windows extracted in the teacher are selected and input into the decoder to perform the reconstruction task. A breakpoint is set at 80 epochs during training, where the parameters of the student encoder are updated to the teacher.
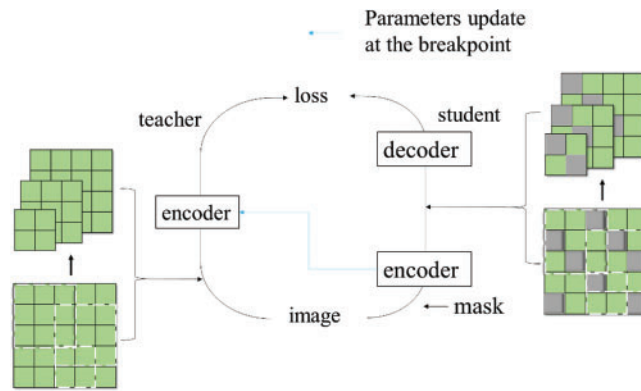
Figure 3: EMLR-seg's schematic diagram. This pre-training model improves the efficiency of general feature learning by reconstructing multi-scale local visual fields. The teacher network does not update gradients during training

3.2 Fine-Tuning

In order to achieve the element segmentation, the pre-trained model needs to be spliced with a segmentation head. In the fine-tuning of the MIM model, UperNet is a commonly used segmentation head. It is a model for image segmentation and has a good understanding of multi-scale semantic information. In order to better compare OSPC-seg's ability with others, we also adopt UperNet as the segmentation head. As shown in Fig. 4, the multi-scale features are input into UperNet after feature transformation, and the segmentation results are output after a series of feature fusions.
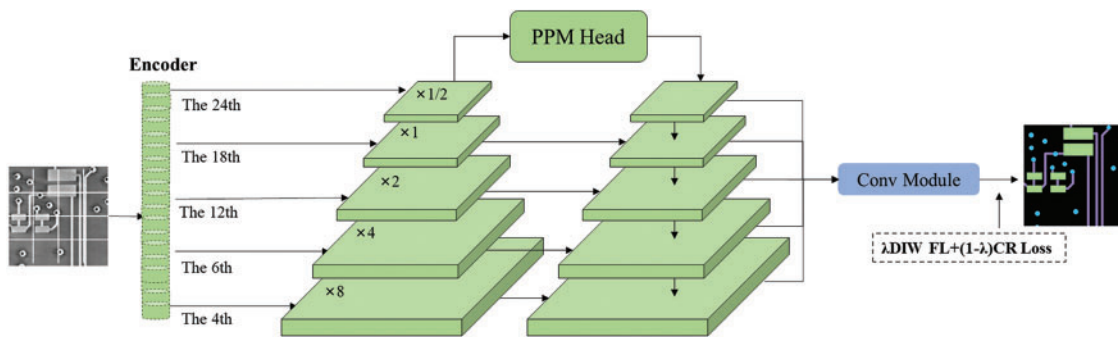

Figure 4: Fine-tuning model schematic diagram. The image is firstly input into the pre-trained encoder, and then the features of the 4th, 6th, 12th, 18th, and 24th layers are dimensionally transformed. The transformed features are output to the feature pyramid. Then a series of feature fusions are performed to output the segmentation results

3.3 OSPC Loss

OSPC Loss includes two sub-loss functions. The Dynamic IoU-Weighted Focal Loss (DIW FL) is used to alleviate the class imbalance problem, and the Connectivity Rate sub-loss function (CR Loss) is used to improve the correct connectivity rate in segmentation results.

The commonly used loss function to solve the class imbalance problem in semantic segmentation is Focal Loss [22], and the formula is:

$$FL = -\alpha \, (1 - p_c)^\gamma \log (p_c) \tag{1}$$

where p_c is the model's estimated probability for the class. α is a hyperparameter used to balance the foreground and background, and γ is a weight about the difficult sample. In order to adapt to multi-class semantic segmentation, we change α in Focal Loss to a four-dimensional tensor α_c, and each dimension corresponds to the weight of the background, pad, wire, and via. According to the explanation in the Focal Loss's paper [22], there is a connection between γ and α, and in the case of $\gamma = 2$, the best result can be achieved when $\alpha = 0.25$. Therefore, we set the minimum value (baseline) of the α_c components to a fixed value to prevent excessive deviation, which will be discussed in the experiment section. In order to dynamically control the weights of each class from the aspect of the evaluation index, we calculate the Intersection and Union (IoU) of each class in current epoch as the α_c component. The formula is as follows:

$$IoU_k = \frac{GT_k \cap Pred_k}{GT_k \cup Pred_k} \tag{2}$$

where k represents the class, GT_k represents the ground truth of class k and $Pred_k$ represents the prediction result of class k. Then α_c is expressed as:

$$\alpha_c = 1 - IoU_k \tag{3}$$

Then adjust the minimum value of α_c to a fixed value, and adjust the other classes in the same proportion. Then the formula of DIW FL is:

$$DIW\_FL = -\,(1 - IoU_k) \, (1 - p_c)^\gamma \log (p_c) \tag{4}$$

To calculate the correct number of connections in the segmentation image, all incorrect connections need to be ruled out. First of all, all the connected regions of the label and the segmentation result are calculated. If a connected area in the label has multiple connected areas at the corresponding position in the segmentation result, it is considered to be a disconnection, recorded as *connect_defect*. Otherwise, it is considered to be an incorrect connection, recorded as *connect_error*. Calculate the ratio of the number of disconnections and incorrect connections to the number of correct connections *target_connect*. The formula is as follows:

$$CR\_Loss = \frac{connect\_defect + connect\_error}{target\_connect} \tag{5}$$

Correct Connectivity Rate (CCR) evaluation index can be expressed as:

$$CCR = 1 - CR\_loss \tag{6}$$

It can be used as an evaluation index to determine the correct connectivity rate in the PCB CT image element segmentation results.

The total loss OSPC Loss is obtained by adding the above two loss functions linearly, and the formula is as follows:

$$OSPC\_Loss = \lambda DIW\_FL + (1 - \lambda) \, CR\_Loss \tag{7}$$

where λ represents impact proportion of the loss function on the model.

4 Experiments

4.1 Datasets

Since there is no public dataset, we created our own PCB CT image datasets. The pre-training dataset includes 400,000 PCB CT images with a resolution of 600 × 600 or above. The fine-tuning dataset consists of sample pairs, each of which contains a CT image and a semantic segmentation label. The training set, test set, and validation set contain 2366 pairs, 500 pairs, and 718 pairs, respectively. We created five validation sets to test the stability of the model. The green, purple, and blue labels correspond to the three elements of pads, wires, and vias, respectively. The example is shown in Fig. 5.
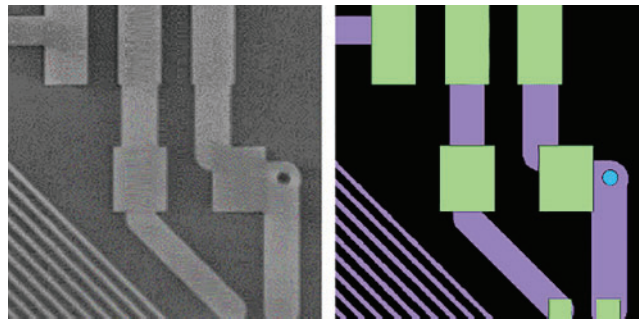

Figure 5: An example of a dataset sample pair. The left image is a CT image, and the right image is a semantic segmentation label

4.2 Default Settings and Indexes

Based on Pytorch architecture, all experiments use distributed data-parallel training on 4 Tesla V100 DGXS by default, with 300 epochs of pre-training and 60 epochs of fine-tuning. In pre-training, random cropping with a scale range of 0.9–1, random horizontal flipping and normalization are used as data enhancements. In fine-tuning, random horizontal flipping and normalization are used as data augmentation. The training parameter settings are as Table 1.

Table 1: AdamW optimizer parameter settings

AdamW optimizer parameter settings	Pre-training	Fine-tuning
Number of warm-up epochs	40	5
Initial learning rate	1.5×10^{-4}	10^{-3}
Weight decay	0.05	0.05

This method uses two evaluation indexes. One is the commonly used semantic segmentation evaluation index mIoU, whose formula is:

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{GT \cap Pred}{GT \cup Pred} \tag{8}$$

where k represents the number of classes. *GT* represents the ground truth and *Pred* represents the prediction result.

The other is the newly proposed index CCR, whose formula is:

$$CCR = 1 - \frac{connect\_defect + connect\_error}{target\_connect} \tag{9}$$

4.3 Comparative Experiment

In the comparative experiment, we used the classic semantic segmentation model for comparison. EMLR-seg was used as the baseline model. All models were tested using exactly the same datasets and parameter settings. As shown in the Table 2, OSPC-seg outperforms the baseline model in terms of both mIoU and CCR, with mIoU reaching 90.1%, exceeding the baseline model by 1.5%, and CCR reaching 97%, exceeding by 1.6%. As can be seen from the visualization images in Fig. 6, OSPC-seg has better segmentation results at the location where elements are connected, which also reflects the improvement of the ability to perceive the semantics of connected relationships. It also reflects the effectiveness of the CCR indicator. When OSPC-seg is not pre-trained, the mIoU is much lower than the value after pre-training, indicating that pre-training on the PCB CT dataset is necessary.

Table 2: Comparative experiments of PCB CT image element segmentation based on commonly used semantic segmentation models. The numbers in the parentheses mean standard deviations

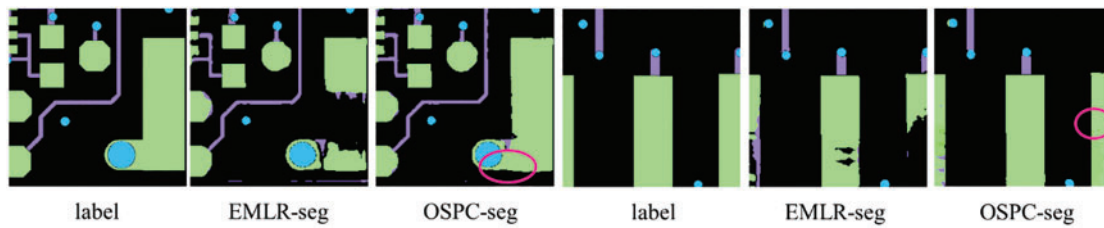|	Model	Pre-training or not	mIoU (%)	CCR (%)
Supervised model	U-Net [25]	×	79.1 (0.2)	\
	U-2-Net [26]	×	84.6 (0.2)	\
	PSPnet [27]	×	82.6 (0.3)	\
	Deeplab v3+ [28]	×	84.3 (0.2)	\
	SETR [29]	×	76.2 (0.1)	\
	Segformer [30]	×	82.3 (0.1)	\
Self-supervised model	MAE [10]	√	86.7 (0.1)	93.9 (0.2)
	CD-MAE [1]	√	87.5 (0.2)	95.1 (0.1)
	dBOT [19]	√	87.4 (0.2)	94.0 (0.2)
	EMLR-seg [2]	√	88.6 (0.2)	95.4 (0.3)
	OSPC-seg	×	77.8 (0.1)	90.0 (0.2)
	OSPC-seg	√	**90.1**$^{\uparrow 1.5}$ (0.2)	**97.0**$^{\uparrow 1.6}$ (0.3)

Figure 6: (Continued)

Figure 6: Visualization of semantic segmentation results. Compared with EMLR-seg, the accuracy of OSPC-seg's connection position recognition is greatly improved

As shown in the Table 3, when OSPC-seg is adopted, the application of OSPC Loss can increase mIoU by 1.4% compared with Focal Loss, which illustrates the effectiveness of OSPC Loss.

Table 3: Comparative experiments of some losses that alleviate class imbalance

Loss	mIoU (%)	Loss	mIoU (%)
CE Loss	88.6	IoU Loss	88.1
WCE Loss	88.4	Lovasz Loss	88.3
Focal Loss	88.7	**OSPC Loss**	**90.1**

We also compared the number of parameters and inference time during fine-tuning using different self-supervised models, and the results are shown in Table 4. Since only the pre-trained encoder is used, the number of parameters is almost the same, but since the loss function of OSPC-seg needs to traverse all pixels to find connected locations, the inference time increases significantly.

Table 4: Comparative experiments of parameters and inference time

Model	Params (M)	Inference time of one epoch (s)
MAE [10]	458.24 M	90
CD-MAE [1]	458.32 M	130
dBOT [19]	458.24 M	120
EMLR-seg [2]	458.24 M	120
OSPC-seg	458.24 M	**250**

4.4 Ablation Experiment

In order to determine whether each sub-loss of the loss function is effective, we conducted ablation experiments on sub-loss functions DIW FL and CR Loss. As shown in Table 5, when only DIW FL is added, mIoU increases by 0.7%, and it can be found from Fig. 7 that the IoU of each element increases by 0%, 0.9%, 0.2% and 0.6%, respectively. The smaller the proportion of pixels, the more IoU is improved, indicating that the problem of class imbalance has been alleviated to a certain extent. When only CR Loss is added, CCR increases by 2.1%, indicating that CR Loss can improve the correct connectivity rate of segmentation result, that is, the perception of the circuit connectivity relationship. When OSPC Loss is used, the CCR index increases by 0.6% compared to only using DIW

FL, indicating that alleviating the class imbalance problem will affect the semantic perception of the connectivity relationship.

Table 5: Ablation experiments of sub-loss functions. The baseline loss is Focal Loss

DIW FL	CR Loss	mIoU (%)	CCR (%)
		88.7	94.3
√		89.4	96.6
	√	88.7	96.4
√	√	**90.1**	**97.0**
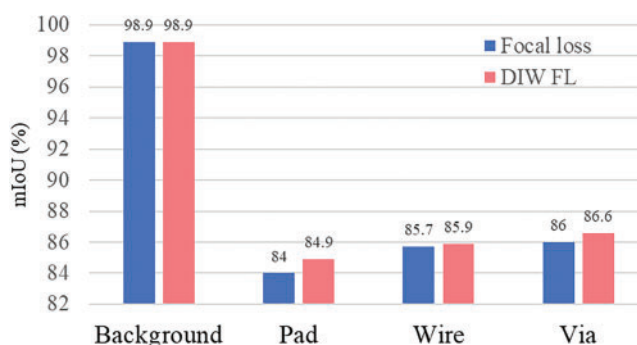
Figure 7: The segmentation effect of different elements using DIW FL and Focal Loss. The pad with the smallest pixel proportion has the most significant improvement

4.5 Hyperparameter Experiment

4.5.1 Baseline of α_c

According to the discussion in the literature [22], there is a connection between γ and α_c, so the ablation experiment should consider both at the same time. We first fixed the value of γ, and then select the baseline value of α_c in the range of 0 to 5. Table 6 shows the optimal effect that can be achieved when γ is determined. The results show that the optimal effect can be achieved when $\gamma = 2$, $\alpha_c = 0.25$. Fig. 8 specifically lists the changes in mIoU when $\gamma = 2$ and baseline of α_c takes different values.

Table 6: Ablation experiments of γ and baseline of α_c. The value range of γ is consistent with that set in [22]

γ	α_c	mIoU (%)
0	0.8	89.0
0.1	0.8	89.1
0.2	0.8	89.3
0.5	0.5	89.3
1	0.25	89.8

(Continued)

Table 6 (continued)

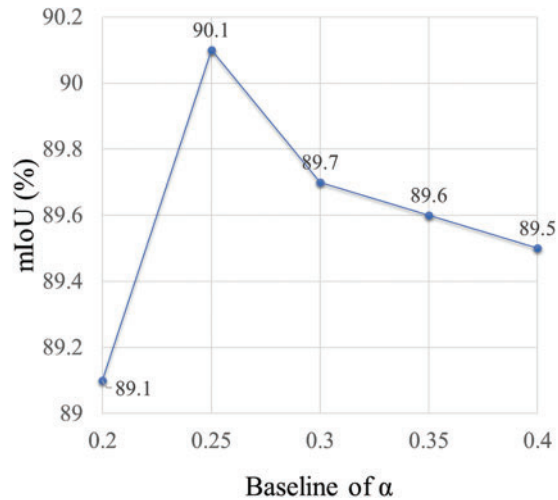| γ	α_c	mIoU (%)
2	**0.25**	**90.1**
5	0.25	89.9

Figure 8: Ablation experiment results of baseline of α_c. The best performance can be achieved when baseline of $\alpha_c = 0.25$

4.5.2 λ

Hyperparameters λ affect the influence degree of the two sub-loss functions on the model. As shown in Fig. 9, When DIW FL accounts for a larger proportion, the segmentation result is more accurate.

4.5.3 Number of Epochs before the Breakpoint

As mentioned in Section 3.1, a breakpoint is set in the pre-training of the literature [2], and the parameters of the teacher network are randomly initialized before the breakpoint. Since the number of training epochs before the breakpoint will affect the performance of pre-training, we conducted experiments on it. The range of the experiment is from 50 epochs to 100 epochs. As shown in the Table 7, when there are 80 epochs before the breakpoint is used, the best transfer effect can be achieved on the segmentation task.
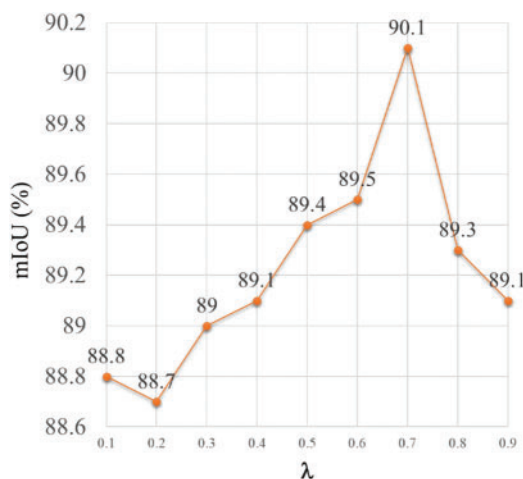
Figure 9: Ablation experiment results of hyperparameters λ. The best performance can be achieved when $\lambda = 0.7$

Table 7: Ablation experiments of number of epochs before the breakpoint

Number of epochs	mIoU (%)
50	88.7
60	88.9
70	89.1
80	**90.1**
90	89.7
100	89.5

5 Conclusions

OSPC-seg adopts a "pre-training and fine-tuning" training process, mainly improving the semantic perception ability of circuit connectivity from the aspect of the loss function. In order to alleviate the problem of class imbalance, a dynamic IoU-weighted focal loss (DIW Focal Loss) is designed. In order to directly improve the correct connectivity rate in the segmentation results, a connectivity rate sub-loss function (CR Loss) is designed. The two are linearly added to obtain the loss that optimizes the semantic perception ability of circuit connectivity (OSPC Loss), and based on the loss, the Correct Connectivity Rate (CCR) evaluation index is proposed as the evaluation standard for connectivity recognition ability. The experiment results show that the CCR of this method on our dataset is improved by 1.6% compared with the baseline model, and mIoU is improved by 1.5%, which demonstrates that the new loss can enhance the model's perception ability of connectivity. The visualization results also confirm that the segmentation accuracy at the connection location is improved. In addition, OSPC-seg considers the class imbalance problem from the aspect of class reweighting. In the future, it can also be considered from the aspect of data resampling to further alleviate its impact on connectivity perception.

References

[1] B. Song *et al.*, "CD-MAE: Contrastive dual-masked autoencoder pre-training model for PCB CT image element segmentation," *Electronics*, vol. 13, no. 6, pp. 1006–1020, 2024. doi: 10.3390/electronics13061006.

[2] C. Chen, K. Qiao, J. Yang, J. Chen, and B. Yan, "Efficient pretraining model based on multi-scale local visual field feature reconstruction for PCB CT image element segmentation," 2024, *arXiv:2405.05745*.

[3] K. Qiao, L. Zeng, J. Chen, J. Hai, and B. Yan, "Wire segmentation for printed circuit board using deep convolutional neural network and graph cut model," *IET Image Process.*, vol. 12, no. 5, pp. 793–800, 2018. doi: 10.1049/iet-ipr.2017.1208.

[4] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Computer Vision–ECCV 2018*. Cham: Springer International Publishing, 2018, pp. 432–448.

[5] T. J. Atherton and D. J. Kerbyson, "Size invariant circle detection," *Image Vis. Comput.*, vol. 17, pp. 795–803, 1999. doi: 10.1016/S0262-8856(98)00160-7.

[6] Y. Yin, H. Luo, J. Sa, and Q. Zhang, "Study and application of improved level set method with prior graph cut in PCB image segmentation," *Circ. World*, vol. 46, pp. 55–64, 2020. doi: 10.1108/CW-03-2019-0028.

[7] D. Li, C. Li, C. Chen, and Z. Zhao, "Semantic segmentation of a printed circuit board for component recognition based on depth images," *Sensors*, vol. 20, no. 18, 2020, Art. no. 5318. doi: 10.3390/s20185318.

[8] D. Makwana, T. R. Sai Chandra, and S. Mittal, "PCBSegClassNet—A lightweight network for segmentation and classification of PCB component," *Expert. Syst. Appl.*, vol. 225, 2023, Art. no. 120029. doi: 10.1016/j.eswa.2023.120029.

[9] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," in *Tenth Int. Conf. Learn. Rep., ICLR 2022*, Apr. 25–29, 2022, pp. 2678–2696.

[10] K. He, X. Chen, S. Xie, Y. Li, P. Dollar and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.

[11] Z. Xie *et al.*, "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9653–9663.

[12] J. Zhou *et al.*, "Image BERT pre-training with online tokenizer," in *ICLR 2022*, 2022, pp. 716–745.

[13] X. Dong *et al.*, "Perceptual codebook for BERT pre-training of vision transformers," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 1, pp. 552–560, 2023. doi: 10.1609/aaai.v37i1.25130.

[14] C. Wei, H. Fan, S. Xie, C. -Y. Wu, A. Yuille and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14668–14678. doi: 10.48550/arXiv.2112.09133.

[15] K. Yi *et al.*, "Masked image modeling with denoising contrast," in *Eleventh Int. Conf. Learn. Rep.*, 2023, pp. 1611–1629.

[16] X. Dong *et al.*, "Bootstrapped masked autoencoders for vision BERT pretraining," in *Computer Vision–ECCV 2022*. Cham: Springer Nature Switzerland, 2022, pp. 247–264.

[17] L. Wei, L. Xie, W. Zhou, H. Li, and Q. Tian, "MVP: Multimodality-guided visual pre-training," in *Computer Vision–ECCV 2022*. Cham: Springer Nature Switzerland, 2022, pp. 337–353.

[18] Z. Hou, F. Sun, Y. -K. Chen, Y. Xie, and S. Kung, "MILAN: Masked image pretraining on language assisted representation," 2022, *arXiv:2208.06049*.

[19] X. Liu, J. Zhou, T. Kong, X. Lin, and R. Ji, "Exploring target representations for masked autoencoders," in *Twelfth Int. Conf. Learn. Rep. (Poster)*, 2024. Accessed: Jan. 01, 2024. [Online]. Available: https://openreview.net/forum?id=xmQMz9OPF5

[20] Y. D. Ma, Q. Liu, and Z. B. Qian, "Automated image segmentation using improved PCNN model based on cross-entropy," in *Proc. 2004 Int. Symp. Intell. Multimed., Video Speech Process.*, 2004, pp. 743–746.

[21] V. Pihur, S. Datta, and S. Datta, "Weighted rank aggregation of cluster validation measures: A monte carlo cross-entropy approach," *Bioinformatics*, vol. 23, pp. 1607–1615, 2007. doi: 10.1093/bioinformatics/btm158.

[22] T. -Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *2017 IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2999–3007.

[23] Y. -F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang and T. Tan, "Focal and efficient IoU loss for accurate bounding box regression," *Neurocomputing*, vol. 506, pp. 146–157, 2022. doi: 10.1016/j.neucom.2022.07.042.

[24] M. Berman, A. R. Triki, and M. B. Blaschko, "The Lovász-Softmax Loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4413–4421.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, Cham: Springer International Publishing, 2015, pp. 234–241.

[26] S. Peng, C. Guo, X. Wu, and L. -J. Deng, "U2Net: A general framework with spatial-spectral-integrated double U-Net for image fusion," in *Proc. 31st ACM Int. Conf. Multimed.*, 2023, pp. 3219–3227.

[27] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6230–6239.

[28] L. -C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[29] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 6877–6886.

[30] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2024, pp. 12077–12090.