electronics

MDPI

Article

CD-MAE: Contrastive Dual-Masked Autoencoder Pre-Training Model for PCB CT Image Element Segmentation

Baojie Song, Jian Chen, Shuhao Shi, Jie Yang, Chen Chen, Kai Qiao * and Bin Yan

Henan Key Laboratory of Imaging and Intelligent Processing, People's Liberation Army (PLA) Strategic, Support Force Information Engineering University, Zhengzhou 450001, China
* Correspondence: qiaokai1992@gmail.com

Abstract: Element detection is an important step in the process of the non-destructive testing of printed circuit boards (PCB) based on computed tomography (CT). Compared with the traditional manual detection method, the image semantic segmentation method based on deep learning greatly improves efficiency and accuracy. However, semantic segmentation models often require a large amount of data for supervised training to generalize better model performance. Unlike natural images, the PCB CT image annotation task is more time-consuming and laborious than the semantic segmentation task. In order to reduce the cost of labeling and improve the ability of the model to utilize unlabeled data, unsupervised pre-training is a very reasonable and necessary choice. The masked image reconstruction model represented by a masked autoencoder is pre-trained on the unlabeled data, learning a strong feature representation ability by recovering the masked image, and shows a good generalization ability in various downstream tasks. In the PCB CT image element segmentation task, considering the characteristics of the image, it is necessary to use a model with strong feature robustness in the pre-training stage to realize the representation learning on a large number of unlabeled PCB CT images. Based on the above purposes, we proposed a contrastive dual-masked autoencoder (CD-MAE) pre-training model, which can learn more robust feature representation on unlabeled PCB CT images. Our experiments show that the CD-MAE outperforms the baseline model and fully supervised models in the PCB CT element segmentation task.

Keywords: unsupervised pre-training; image segmentation; PCB nondestructive testing; model finetuning

Citation: Song, B.; Chen, J.; Shi, S.; Yang, J.; Chen, C.; Qiao, K.; Yan, B. CD-MAE: Contrastive Dual-Masked Autoencoder Pre-Training Model for PCB CT Image Element Segmentation. *Electronics* **2024**, *13*, 1006. https://doi.org/10.3390/electronics13061006

Academic Editor: Byung-Gyu Kim

Received: 1 December 2023
Revised: 28 December 2023
Accepted: 3 January 2024
Published: 7 March 2024

1. Introduction

Printed circuit boards are a core part of electronic devices by connecting various types of components to achieve various functions. For some important and complex electronic equipment, regular inspection and maintenance are required, where printed circuit board testing technology based on computed tomography provides a nondestructive inspection method, such as [1,2]. Regarding the different elements on the printed circuit board, the vias provide space for the arrangement of components, and the pads and wires provide pathways for the connection of components, so the detection of wires, vias, and pads is the key to the entire nondestructive testing process. With the continuous development in the field of computer vision, some algorithms based on deep neural networks have been introduced for element detection, such as the use of image semantic segmentation technology to achieve the segmentation of the wires [3] and the detection of vias [4]. Compared with the traditional manual detection methods, deep learning-based image segmentation methods can be very effective in improving efficiency and accuracy. However, deep learning models often require a large amount of annotated data to perform well, which will result in huge labeling costs.

In order to adequately train the model and extract potential features on limited labeled data, pre-training is a very effective and necessary method. Based on this, by building an efficient pre-training model, a powerful feature extractor is obtained after learning on a large

amount of unlabeled data, and then the pre-trained model is fine-tuned on a small amount of labeled data to perform the PCB CT image element segmentation task. This approach is important to reduce the dependence of the model on labeled data and reduce the labeling cost. For the above purposes, unsupervised pre-training is a very good solution. Unlike supervised pre-training, unsupervised pre-training does not require the supervision of labeled data but learns feature representation on unlabeled datasets by self-supervised or unsupervised means. In addition, the success of large language models such as Transformer [5] has inspired researchers in the field of computer vision. As a higher-dimensional modality with more noise and more redundant information than text, modeling on images is more difficult. However, the proposal of Vision Transformer [6] shows great potential of transformer in computer vision tasks, as shown in some remarkable works such as BEiT [7], iGPT [8], MAE [9], etc. These transformer-based models with a self-attentive mechanism have a significant advantage of being able to work on a large amount of unlabeled data, in addition to having good scalability and the ability to obtain global information.

Unsupervised pre-training is continuously showing great potential, and is mainly divided into two types, generative pre-training and contrastive pre-training, each of which mines its own supervised information from large-scale unlabeled data through different auxiliary tasks. Among them, the agent task of generative pre-training is mainly masked image reconstruction, and the main process is to randomly mask some regions of the original image and then send it to the model for image reconstruction. The operation of masking can increase the learning difficulty for the model, so that the model can learn better feature representation, typical methods are BEiT [7], MAE [9], SimMIM [10], etc. While the agent task of contrastive pre-training is mainly individual discrimination, and the core idea is to reduce the distance between positive samples and to increase the distance between different negative samples by contrast learning, and the main methods are MoCo [11], SimCLR [12], SwAV [13], BYOL [14], DINO [15], etc. All these pre-training methods even outperform the supervised pre-trained models in some downstream tasks, such as target detection and semantic segmentation tasks.

For PCB CT images, the contrastive pre-training method is difficult to be effective because of the small number of element categories, the basically fixed size and the single image containing almost all categories of elements, so we believe that generative pre-training is more suitable for our task. At the same time, due to the limitations of the cone-beam CT imaging method, PCB CT images are of low quality, there exist some phenomena such as metal artifacts, grayscale non-uniformity and translucency, as shown in Figure 1, where metal artifacts refer to deformation or distortion around some elements, grayscale non-uniformity refers to slight differences in the grayscale values of the same elements, and translucency refers to the occurrence of results between different layers of the printed circuit board. The appearance of these problems tends to limit the learning efficiency of the model. Therefore, our pre-trained model needs to have better robustness and be able to learn more core features in the PCB CT images. Through the above analysis, we propose a contrastive dual-masked PCB CT images pre-training method, which adds the idea of contrast learning to the original masked autoencoder [9], and can enable the model to learn more robust feature representation capabilities by pulling closer the features obtained by encoding the same image after masking in different regions. The specific implementation details will be presented in Section 3.

Our proposed method introduces the idea of contrastive learning to the process of masked image reconstruction. Compared with the contrastive pre-training method, We still use image reconstruction as the implementation of pre-training, because for PCB CT images, the difference between positive and negative samples is small, it is difficult to learn effective features simply by the operation of pulling the positive samples closer and pulling the negative samples farther, while the model can be ensured to learn relatively more useful feature information by means of image reconstruction. The comparative experiments can demonstrate that our pre-training method extracts features more robustly for PCB CT images and achieves a very satisfactory performance in downstream tasks.

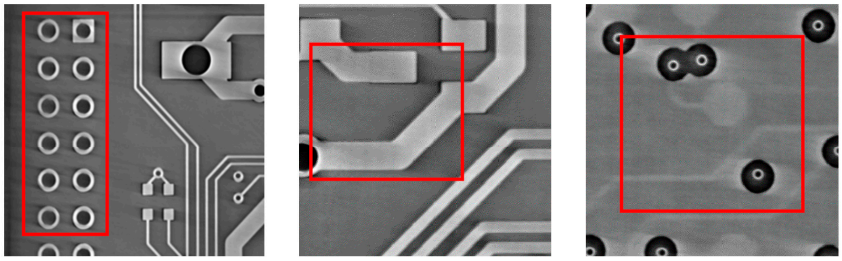

Figure 1. PCB CT images with low image quality.

To summarize, our main contributions are:

1. We proposed a PCB CT image pre-training method based on contrastive dual-masked image reconstruction, in which the features obtained by masking different regions of the same image are pulled closer to improve the robustness of the features.
2. We introduced the paradigm of "Pre-training & Fine-tuning" to the PCB CT image element segmentation task, in which the model is first pre-trained on a large amount of unlabeled data to make the model have good feature representation capability, and then the pre-trained model is supervised and fine-tuned on a small amount of labeled data to achieve the final element segmentation downstream task. This paradigm can effectively reduce the labeling cost and shows great performance in downstream element segmentation task.
3. We experimentally demonstrated that our approach outperforms original MAE and outperforms fully supervised models based on CNN architecture.

2. Related Works

Unsupervised Pre-training has developed rapidly in computer vision, and a number of methods are now showing powerful capabilities in various tasks and even surpassing some supervised training models. Unsupervised pre-training in computer vision mainly has two approaches, where generative pre-training mainly refers to masked image modeling, such as MAE [9], BEiT [7], etc. Inspired by the field of natural language processing, these methods divide the image into equal-sized patches and analogize them to words in a sentence, by masking random parts of the input patches and then reconstructing the missing pixels to achieve the purpose of understanding the image. A certain percentage of masking increases the learning difficulty but also effectively improves the performance of the model in downstream tasks. In addition to generative pre-training methods, contrastive pre-training models also play a very important role in this field, which focus on learning the common features among similar samples and distinguishing the differences between non-similar samples. Unlike generative pre-training methods, contrastive pre-training does not need to focus on the very tedious details of the samples, but simply learns the features that will distinguish them from other samples. Examples include InsDisc [16] and InvaSpread [17], which used individual discrimination tasks early as proxy tasks, and CPC [18], which used generative contrastive learning to learn the future output or other negative samples in contrast to the predicted output, and CMC [19], which increased mutual information by increasing different perspectives. With further research, some studies found that the size of the batch in the training process has a great impact on the performance of the contrastive pre-training model, so based on InvaSpread, Chen et al. [12] proposed SimCLR, which treats the augmented data as positive samples and all other images as negative samples, and it means that a positive sample needs to be matched with multiple negative samples, otherwise it is difficult for the model to converge. To address this problem, He et al. proposed the MoCo [11] based on the momentum approach. According to the further improvement of these models, SimCLR v2 [20], MoCo v2 [21] and MoCo v3 [22] were subsequently proposed.

Furthermore, some methods have emerged that do not require negative samples, such as SwAV [13] which combines contrastive learning and clustering methods and does not directly perform the comparison between two samples, but first clusters the samples and then performs contrastive learning between classes. In addition, there are a series of methods such as BYOL [14], SimSiam [23], Barlow Twins [24], and DINO [15]. In summary, there is no doubt that both generative and contrastive pre-training methods have contributed to the further development of the computer vision.

In the PCB CT image element segmentation task, for the characteristics of the CT images of printed circuit boards, we believe that the generative pre-training method is simpler to realize and more effective, Therefore, we chose the masked auto-encoder as our baseline and proposed an improved contrastive dual-masked autoencoder pre-training model for PCB CT Images.

Image Segmentation is one of the mainstream tasks in the field of computer vision, which is widely used in many aspects such as intelligent keying, autonomous driving, medical image diagnosis and human–computer interaction. As more and more application scenarios require accurate and efficient segmentation techniques, image semantic segmentation has received more attention and importance. Image segmentation can be considered as a pixel-level understanding of an image, which is essentially the classification of each pixel point in an image and ultimately the representation of objects of a certain class using the same class label. The full convolutional network [25], as the pioneer of deep learning in image semantic segmentation, also laid down the basic structural form of encoder–decoder for image segmentation models. After that, a series of improved methods based on FCNs were born. For example, the U-Net [26] with the addition of skip connection, which differs from the fusion operation of FCN summation, uses a concatenation approach to fuse deep and shallow features. This model can achieve better segmentation results with little training data and is widely used in medical image segmentation tasks. Qin et al. [27] proposed a two-layer nested U-shaped model based on the U-net, which can train the model from scratch without relying on the pre-trained model, and the feature extraction is as good as the pre-trained model.

In addition to this, attention has started to be paid to the importance of contextual information for image understanding, due to the limitation of the convolutional structure, the receptive field of models based on CNNs is always limited and therefore cannot make good use of global contextual information. In order to overcome this problem and enable the model to focus on the global information of the input image more effectively, some methods based on dilated convolution and image pyramid structure are proposed, such as PSPNet [28], Deeplabv3+ [29], etc. Moreover, for allowing the model to highlight certain important features of the object, some methods incorporate attention mechanisms to ignore irrelevant information and better focus on the key information, such as DANet [30], EMANet [31], etc. Besides the above methods, there are many other CNN-based image semantic segmentation methods that show satisfactory results.

With good scalability and powerful global information acquisition capability, transformers based on the self-attention mechanism have started to gain much attention in the field of computer vision. With the introduction of vision transformer [6], transformer-based methods began to gradually replace CNNs as the main architecture for various vision tasks. For image segmentation tasks, some transformer-based methods [32–34] are proposed and showed a good result. Recently, some basic large models dedicated to image segmentation have started to appear and attract much attention. For example, SAM [35] proposed by Meta is a model based on a massive data training of 11 million images and 1.1 billion masks and has a strong zero-shot performance to segment unseen images very effectively. Importantly, the model pioneers the combination of image segmentation and prompt which include points, boxes, and text, so it can realize the segmentation of different images with the help of prompt. After that, there are similar works such as SegGPT [36] proposed by BAAI and SEEM [37] proposed by Microsoft. The presentation of

these models also opens a new era of generalized large models in computer vision, and further promotes the rapid advancement of the image segmentation.

There are various methods for image semantic segmentation, but for the difference between PCB CT images and natural images, we needed to choose a suitable method for our task, so we have conducted experiments using several typical methods, such as U-Net, PSPNet, EMANet, Deeplabv3+, etc. The specific experimental results are shown in Table 1. Although these methods can achieve end-to-end element segmentation, the results are not very satisfactory. These models need more labeled data to perform better, which will lead to a large training cost. In view of this, we first apply the "Pre-training & Fine-tuning" paradigm to PCB CT element segmentation task, and verify the effectiveness of our method through experiments, which can reduce the reliance of the model on labeled data, improve the utilization of unlabeled data, and accelerate the training and deployment of the model.

Table 1. Performance comparison between CD-MAE and other methods on PCB CT image element segmentation task.

Method	Pre-Trained	Backbone	MIoU(%)	#Param.
SegNet [38]		—	77.8	29.45 M
U-Net		—	79.1	31.04 M
U-2-Net		—	84.6	44.05 M
PSPNet		ResNet152	82.6	71.44 M
EMANet		ResNet152	79.8	69.41 M
Deeplabv3+		ResNet152	84.3	72.33 M
SETR		ViT-L	76.1	317.3 M
SegFormer		SegFormer-B4	82.2	60.89 M
MAE		ViT-L	77.6	458.24 M
MAE	√	ViT-L	86.5	458.24 M
SimMIM	√	Swin-L [39]	78.5	175.52 M
CD-MAE	√	ViT-B	86.9	172.98 M
CD-MAE	√	ViT-L	87.5	458.32 M
CD-MAE	√	ViT-H	88.3	872.98 M

PCB Nondestructive Testing refers to the use of cone beam CT to image the printed circuit boards in electronic equipment, and then use certain technical means to analyze and diagnose the PCB CT images to achieve the purpose of nondestructive maintenance and analysis of important electronic equipment. Since the elements of printed circuit boards such as vias, pads, and wires connect various components and thus achieve different functions, the inspection of these elements is a key step in the whole process. With the development of deep learning, some element detection methods based on deep neural networks provided a new approach for element detection, such as the wire segmentation method using deep convolutional neural networks combined with graphical cut models [3], the component segmentation method [40] that constructs a random forest pixel classifier, and the use of Mask R-CNN [41] implementation of vias detection [4]. These methods have not been very maturely applied in the actual nondestructive testing process, but they have validated the great potential of image semantic segmentation based on deep learning in element detection.

We made many attempts to better promote the application of image semantic segmentation in the element detection task. Aiming at the characteristics of PCB CT images and the cost of data annotation, we finally adopted the route of unsupervised pre-training before supervised fine-tuning. We first put CD-MAE on a large amount of unlabeled data to adequately learn the feature representation of elements in PCB CT images. After the pre-training, we select the encoder as the feature extraction module and adopt UperNet [42] as the segmentation head, and then the whole segmentation network is fine-tuned on a small amount of labeled data to finally achieve PCB CT image element segmentation, where the

parameters of the pre-trained encoder part will be frozen in this stage. Through experiments, we demonstrate that our pre-training method outperforms MAE and surpasses some classical supervised models.

3. Method

Our proposed pre-training method for PCB CT images is an improvement on the masked autoencoder model. So, as with the original MAE, we continue to follow the original basic steps of masking, encoding and decoding. However, the difference is that our method incorporates contrast learning, that is, we first perform two random masking operations on the same sample in different regions and apply a pulling operation to the encoded features. In the whole process, image reconstruction can ensure the effectiveness of feature extraction, while the feature pulling operation can ensure the robustness of feature extraction.

3.1. Network Structure

The structure of our method is shown in Figure 2, which mainly consists of an image masking part, a parameter-shared encoder, and a reconstruction decoder. CD-MAE has no major difference in the process with MAE, which basically follows the operation of masking the image randomly, encoding the non-masked patches, and adding the masked tokens for image reconstruction. The biggest difference is that our proposed method is to mask different regions of the same image and then pass the visible patches through a parameter-shared encoder, respectively to obtain the features, and then the two groups of features are compared and pulled closer to ensure robustness to feature extraction.

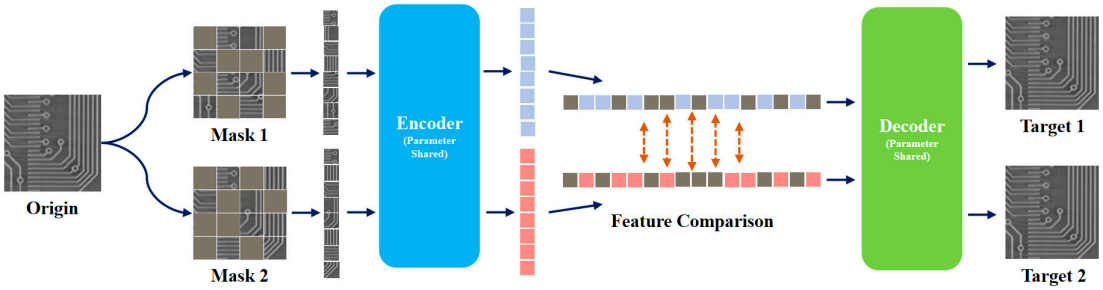

Figure 2. Overview of our proposed CD-MAE.

The first step in masked image modeling is to randomly mask a certain ratio of the image patches, and as explained in the paper of masked autoencoder [9], the model still performs well after randomly masking, probably because some degree of masking largely eliminates information redundancy while a highly sparse input also helps train an efficient encoder. At the same time, the masking operation makes the image reconstruction difficult and allows the model to learn the higher dimensional feature representation of the image rather than staying at the underlying information such as pixels. Different from the usual masking processing, CD-MAE not only masks a certain percentage of image patches, but also performs the random masking of the same image twice in different regions. This operation is designed to be able to subsequently compare the features obtained from different areas of the same image.

The encoder still uses ViT, and the input is the unmasked image patches. After masking the same image twice in different regions, the two groups of visible image patches are used as input, through linear projection, adding positional embeddings, and then a series of transformer blocks with shared parameters to, respectively obtain the corresponding intermediate features. We believe that for the same PCB CT image, the semantic information in the image should be fixed due to the relatively fixed elements and the connection relationship between them, so the features obtained by the encoder should be consistent even if different regions are masked. Such operation will allow the model to learn a more robust and core semantic representation. In order to optimize the encoder towards this

goal, we use the mean squared error (MSE) loss to pull the two groups of features closer together, as shown in Equation (1).

$$L_{FC} = MSELoss(g(f(Patch_{UM1})), g(f(Patch_{UM2}))) \tag{1}$$

where L_{FC} represents the loss of the contrast learning loss between the two sets of features, and $Patch_{UMi}$ $(i = 1, 2)$ represents the visible image patches obtained by masking different areas of the same image. $f(\cdot)$ indicates that patches are passed through the encoder, while $g(\cdot)$ indicates adding masked tokens to the encoded visual image patches to restore to the length of the original sequence.

The proxy task of masked image modeling is image reconstruction, so after encoding the image patches, it is also necessary to decode and reconstruct the features. In this part, we feed the entire sequence of image patches with mask tokens to the decoder for reconstruction. The loss in this part is calculated between the reconstructed image and the original image using mean squared error loss. The details are as in Equation (2).

$$L_{Recon_i} = MSELoss(I_{Org}, I_{\mathrm{Recon}_i}) \qquad i = 1, 2 \tag{2}$$

I_{Org} represents the original image, while I_{Recon_i} represents the reconstructed image.

The loss function of the whole network is shown in Equation (3), where λ is a hyperparameter that represents the weight of the loss.

$$L = L_{\mathrm{Recon}_1} + L_{\mathrm{Recon}_2} + \lambda * L_{FC} \tag{3}$$

3.2. Pre-Training

The purpose of the pre-training is to adequately model on a large-scale dataset to obtain a generic model, and subsequently fine-tune or migrate it in different downstream tasks to accomplish different target tasks. For the PCB CT image element segmentation task, although our downstream task is relatively single, we believe that unsupervised pre-training on a large amount of unlabeled data is also necessary considering the cost of data labeling. Therefore, our proposed model, CD-MAE, is fully pre-trained on unlabeled PCB CT images to obtain an encoder with excellent feature representation capability. The encoder can learn deeper features of the image with better robustness. Figure 3 shows the image reconstruction results of our method and MAE with different mask ratios after pre-training.

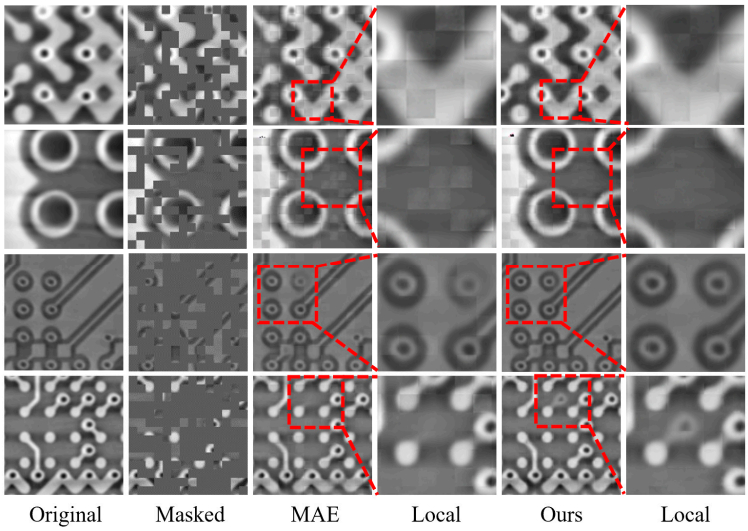

Figure 3. The results recovered by the model after pre-training with different methods. Where the first column is the original PCB CT image, and the second column is the masked image, in which the first two rows have a mask ratio of 0.5 and the last two rows have a mask ratio of 0.7.

From the global view of the image, the two methods have little difference in reconstruction ability. In local and detail terms, the reconstructed images from CD-MAE are a bit finer. For example, we can see from the first two rows of the reconstructed image that the gray scale of images reconstructed by MAE will have some differences in adjacent patches, so we can observe obvious patch shapes in the reconstructed images, while ours is more uniform. From the last two rows of the reconstructed image, it can be concluded that our method can learn deeper semantic information under the condition of a high masking ratio. Even though most of the image is covered, the model pre-trained by CD-MAE can still recover some invisible patches. All of the above are benefits of using dual-masked contrastive learning that the features are more robust. However, the performance of the model on downstream tasks cannot be seen from the reconstruction results alone, so fine-tuning is also needed to verify the effectiveness of the pre-trained model.

3.3. Fine-Tuning

After pre-training, we can obtain an encoder with strong feature representation. According to the classical "encoder–decoder" structure of image semantic segmentation model, the decoder of pre-trained model needs to be replaced by a segmentation head. After replacement, the overall network is trained with supervised fine-tuning on the labeled data to achieve the downstream element segmentation task. At present, there are several well-performed semantic segmentation networks. Here we chose UperNet [39], a unified perceptual parsing network for scene understanding, which is based on the traditional convolutional network architecture and is designed with the idea that the network parses visual concepts at different perceptual levels, such as scene, object, texture, and material, all at once. The principle of UperNet is to construct a feature pyramid network by using the feature map output from the last layer of the feature extraction network to extract multi-scale feature information and use it for target identification and localization at different levels. Most importantly, it performs better in semantic segmentation tasks due to the use of multiple semantic levels of features.

As shown in Figure 4, we combine the pre-trained encoder with the part of UperNet used for segmentation to construct the whole element segmentation network. For the training of the overall network, we freeze all parameters of the pre-trained encoder and only train the segmentation head with cross-entropy loss as in Equation (4).

$$L_{seg} = CrossEntropyLoss(y_{pred}, y_{gt}) = \frac{1}{N}\sum_{i} L_i = -\frac{1}{N}\sum_{i}\sum_{c=1}^{K} y_{ic}\log(p_{ic}) \quad (4)$$

y_{pred} represents the segmentation results of the model, y_{gt} represents the ground truth. N represents the number of pixels, K represents the number of categories, p_{ic} indicates that the model predicts pixels i as categories c. y_{ic} is a symbolic function, taking 1 if the true category i is equal to c, and 0 otherwise.

The pre-trained weights are used for encoder, and the decoder is replaced with UperNet for element segmentation. For the encoder using ViT-L as the backbone, the outputs of layers 6, 12, 18, and 24 are selected as the multi-scale input of UperNet. Then, the final element segmentation results are generated through a series of operations such as pyramid pooling module, feature pyramid network, and feature fusion. Where B is the batch size, H and W are the dimensions of the image, p is the size of the image patch, and C is the number of channels.

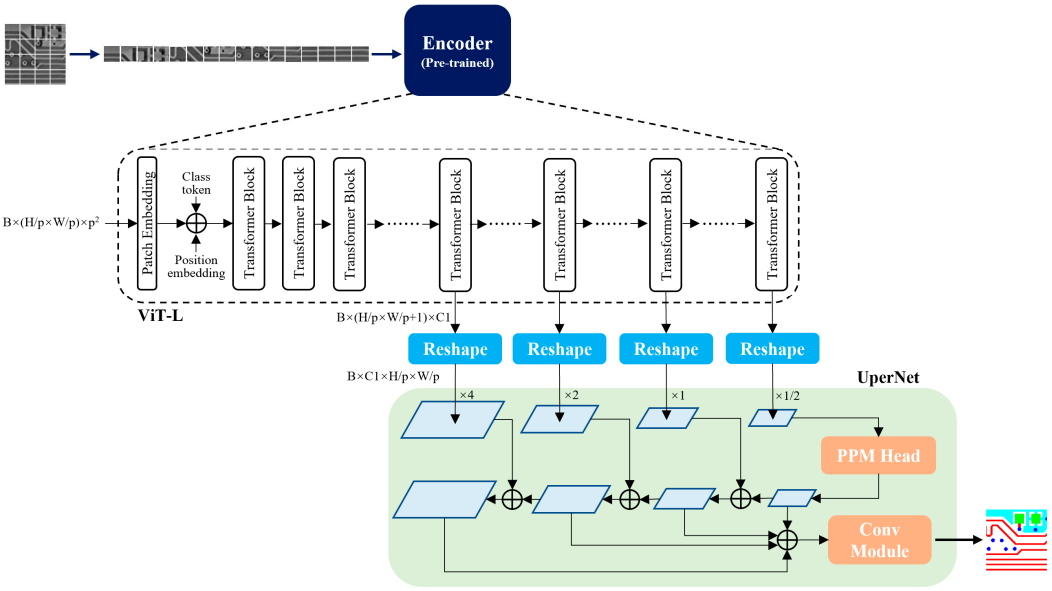

Figure 4. Structure of element segmentation network based on pre-trained model.

4. Experiment

4.1. Preliminaries

Evaluation Metrics. In image semantic segmentation tasks, a commonly used model performance evaluation metric is the intersection over union (*IoU*), which is calculated between the ground truth and the prediction results by dividing the overlapping region of the same category by their union, so it can be used to evaluate the similarity between the segmentation result and the ground truth of a category. In this case, the intersection over union for a single category is calculated as in Equation (5).

$$IoU = \frac{GT \cap Pred}{GT \cup Pred} = \frac{TP}{TP + FP + FN} \tag{5}$$

where $GT \cap Pred$ represents the intersection between the ground truth and the prediction result, and $GT \cup Pred$ represents the union of them. TP denotes the probability of correctly predicting a positive sample, FP denotes the probability of incorrectly predicting a positive sample, and FN denotes the probability of incorrectly predicting a negative sample.

To obtain the segmentation effect for all elements, the mean intersection over union (*MIoU*) is obtained by summing the *IoU* of each category in the dataset and averaging them to represent the predictive effect of the model for all categories.

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{TP + FP + FN} = \frac{1}{k+1}\sum_{i=0}^{k} IoU \tag{6}$$

where k represents the number of categories.

Dataset. The number of unlabeled data in our PCB CT image dataset reaches 400,000 with a resolution of 500 × 500. The total number of labeled datasets is 3584 with a resolution of 500 × 500, of which there are 2366 samples in the training set, 718 samples in the validation set, and 500 samples in the test set. The dataset contains mainly 3 categories of elements, which are pads, wires and vias. Some of the data samples are shown in Figure 5. The unlabeled dataset is used for the training of CD-MAE in the pre-training phase, and the labeled dataset is used for fine-tuning training of the element segmentation network based on the pre-training encoder.

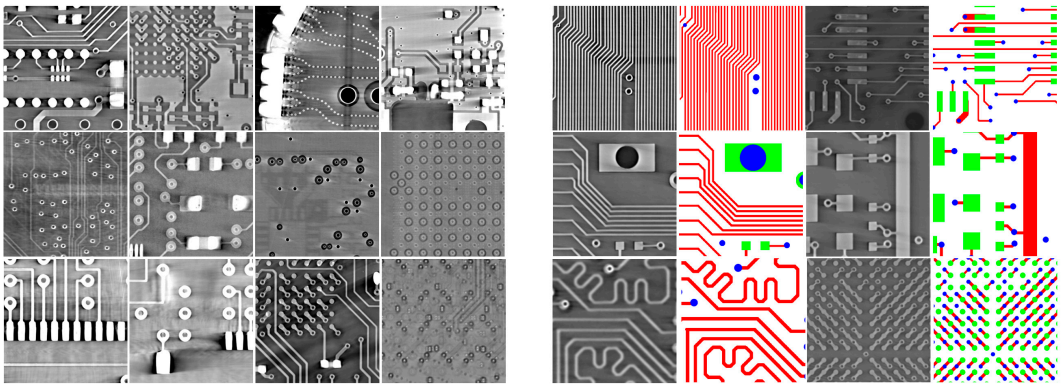

Figure 5. Partial unlabeled data and labeled data.

The left half contains some unlabeled data. The right half contains labeled data pairs, where two columns are in a tuple, one on the left for the original image and one on the right for the ground truth. In the ground truth, red represents the wire, green represents the pad, and blue represents the vias.

4.2. Results on PCB CT Image Element Segmentation

Experimental Setup. In the pre-training phase of our proposed model, we adopted ViT-L as the backbone, which contains 24 transformer blocks of size 1024, and the decoder part consisting of eight transformer blocks of size 512 and a linear prediction layer. The input size of the image is 224 × 224 and the patch size is 16 × 16. Our experiments were conducted on 4 Tesla V100 DGXS with the pre-training epoch set to 100 and the batch size to 64. We employed an AdamW optimizer, and a cosine learning rate scheduler with a 40 epoch warm-up, where the base learning rate was set to 10^{-3}, the weight decay was 0.05, and the weight of feature contrast loss was set to 0.25. In the basic experiment, we used 100,000 unlabeled data for pre-training, in which the data enhancement strategy used random resize cropping with a scale range of [0.9, 1], and random horizontal flip and normalization steps.

In the fine-tuning stage, we replaced the decoder part with the UperNet to achieve the downstream element segmentation task. The batch size of the fine-tuning stage was eight, and the number of fine-tuning epochs was 40. And we still used the AdamW optimizer with a cosine learning rate scheduler, where the warm-up was five epochs, the initial learning rate was 10^{-3}, and the weight decay was 0.05. In order to obtain the input of UperNet at different scales, we selected the outputs of six, twelve, eighteen, and twenty four transformer blocks in the encoder and sent them to UperNet after reshaping and up-sampling at different scales.

Different image masking ratios. In order to test the effectiveness of our proposed pre-training method in the PCB CT image element segmentation task, we conducted comparative experiments with MAE at different mask ratios, and the experiment results are shown in Figure 6, from which we can see that the unsupervised pre-training approach based on masked image reconstruction can indeed play a certain role in element segmentation by using a large amount of unlabeled data. In the pre-training stage, the information redundancy is largely eliminated by randomly masking the image patches at a certain ratio, thus making the image reconstruction difficult, so that the model learns higher dimensional feature expressions of the data instead of staying at the underlying information such as pixels, the specific effect of which will be further explained in Section 4.3. The experimental results also show that a reasonably high mask ratio can increase the training difficulty of the model. Although the model in a low mask ratio can also learn certain knowledge, the effect is far less than that in a high mask ratio. It is obvious from the experiment results that the model pre-trained by CD-MAE performs better than MAE in the element segmentation task at any mask ratio, and all of them exceed the best performance of MAE in the mask

ratio range of 0.4 to 0.8, which can also illustrate the superior robustness of our proposed pre-training method for feature extraction.

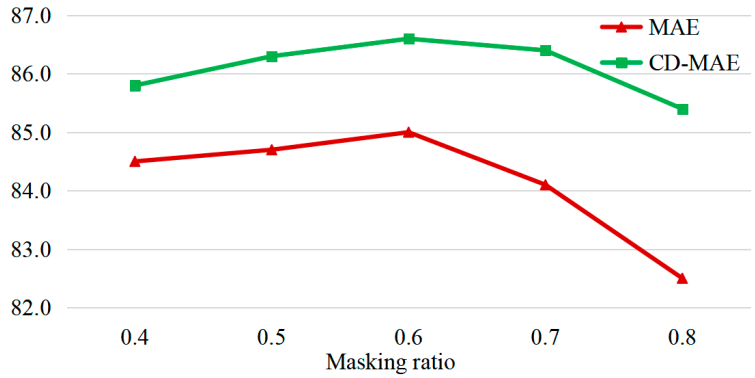

Figure 6. Comparison of fine-tuning results of CD-MAE and MAE with different masking ratios.

Different sizes of pre-training datasets. To further validate the effectiveness of our proposed method, we conducted further experiments, such as changing the dataset size, changing the model size, and comparing with other methods including supervised models. Figure 7 shows the experiments conducted under different dataset sizes, and it can be seen from the experimental results that our method always outperforms MAE with varying dataset sizes. In particular, the effect is more obvious with a small amount of unlabeled data, and the performance of our pre-training method with 50 k unlabeled data already exceeds the performance of MAE with 100 k unlabeled data. However, it is worth noting that with more data, the performance improvement of both MAE and CD-MAE starts to slow down, and the gap between them decreases to a certain extent. We speculate that this is mainly due to the characteristics of PCB CT images, including the small number of classes and the relatively fixed size and shape, thus in the case of particularly large amounts of data, the model is already proficient in the feature information of the elements in PCB CT images, so that adding additional data does not cause too significant a change in performance.

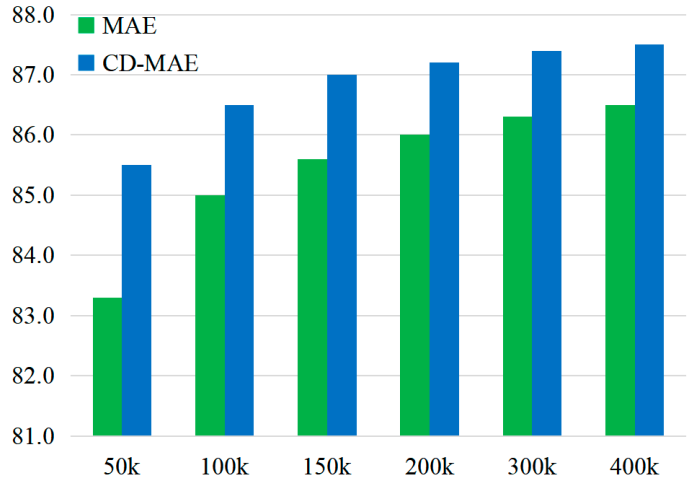

Figure 7. Influence of dataset size on pre-training effect.

Visualization of the segmentation results. We show some of the prediction results from the model based on the encoder pre-trained by CD-MAE in Figure 8. As with the original labeled data, the different colors in the segmentation results represent different PCB elements, for example, red represents wires, green represents pads, and blue represents vias. The element segmentation model based on our proposed pre-training method

has better results for all types of elements in PCB CT images, both in terms of the overall and details such as edges and connections of each element.

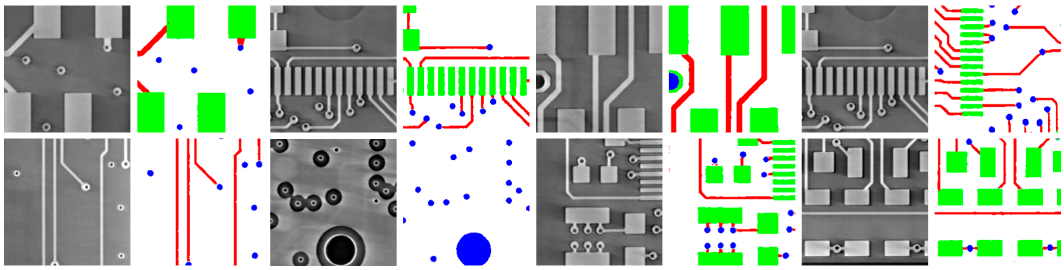

Figure 8. Visualization results of the element segmentation.

4.3. Comparative Experiments

At the same time, we also conducted experiments on the size of the backbone, including the application of ViT-B, ViT-L, and ViT-H models on 400 k unlabeled data, and the results are shown in the last three rows of Table 1. The differences between the three models mainly lie in the number of layers of the transformer block, the embedding dimension, and the size of the patches. It can be seen that the performance of the pre-trained model in the downstream task is further improved as the model size increases, but the improvement is not particularly obvious, which is also related to the characteristics of PCB CT images as we analyzed. In Table 1, we also conducted experiments with other models, which include some CNN-based and transformer-based models. In terms of experiment consistency, for methods that also require unsupervised pre-training, we ensured that the models were all performed on a 400 k unlabeled dataset and used the same segmentation head for fine-tuning. For both the supervised methods and methods based on pre-trained models, the labeled data we used were identical. From the results, we can see that some methods with simple structures can achieve good results in PCB CT image element segmentation tasks, such as U-2-Net, DeepLabv3+, etc. In contrast, the results of some methods with complex structures that perform well on natural images are rather poor. Most importantly, the model pre-trained by CD-MAE significantly outperforms other methods, which fully illustrates the effectiveness of our proposed method and also shows that our application of unsupervised pre-training to the PCB CT image element segmentation task is very effective and necessary.

4.4. Discussion and Further Work

Masked image modeling, as a training method in unsupervised pre-training, relies on the ambiguity brought by the mask. Moreover, for PCB CT images, the elements are small and numerous, so masking the images by a certain ratio can produce serious semantic ambiguity, which means that the model trained with a high masking ratio will misidentify when two categories are visually close, and when one category is much more dominant than the other. From the experimental results, we argue that it is this semantic ambiguity that forces the feature extraction network to learn high-level semantic information in images. This high-level semantic information may be the rules of arrangement of the elements in the PCB, high-level features, or others. Therefore, a certain semantic ambiguity is the key to improving the generalization performance of the model. Regarding the paradigm of "Pre-training & Fine-tuning" on the element segmentation task, our subsequent work will also focus more on fine-tuning, so that the pre-trained models can play a better role in the downstream task and further improve the performance of the model after fine-tuning.

5. Conclusions

In this paper, we propose an improved contrastive dual-masked pre-training model based on MAE, which can improve the robustness of feature extraction by narrowing the distance between features in different mask regions of the same image, and thus play a

better role in the PCB CT images element segmentation task. Moreover, our model pre-trained on a large amount of unlabeled data performs significantly better than the purely supervised training model in the downstream task, which further demonstrates the effectiveness and necessity of pre-training in elements segmentation task.

Author Contributions: B.S. and K.Q. put forward the corresponding ideas and methods. J.C., B.S. and K.Q. wrote the main manuscript text. C.C., J.Y. and S.S. participated in the discussion and validation of the method. B.Y. and J.C. were responsible for the co-ordination and supervision of the whole process. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The PCB CT image data used in current study are all scanned and annotated by us. Because the printed circuit boards involve some intellectual property rights and trade secrets, the datasets are not publicly available but are available from the corresponding author on reasonable request.

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

References

1. Asadizanjani, N.; Shahbazmohamadi, S.; Tehranipoor, M.; Forte, D. Non-destructive PCB reverse engineering using X-ray micro computed tomography. In Proceedings of the 41st International Symposium for Testing and Failure Analysis 2015, Portland, OR, USA, 1–5 November 2015; ASM International: Almere, The Netherlands, 2015; pp. 164–172.
2. Asadizanjani, N.; Tehranipoor, M.; Forte, D. PCB reverse engineering using nondestructive X-ray tomography and advanced image processing. *IEEE Trans. Compon. Packag. Manuf. Technol.* **2017**, *7*, 292–299. [CrossRef]
3. Qiao, K.; Zeng, L.; Chen, J.; Hai, J.; Yan, B. Wire segmentation for printed circuit board using deep convolutional neural network and graph cut model. *IET Image Process.* **2018**, *12*, 793–800. [CrossRef]
4. Botero, U.J.; Koblah, D.; Capecci, D.E.; Ganji, F.; Asadizanjani, N.; Woodard, D.L.; Forte, D. Automated via detection for PCB reverse engineering. In Proceedings of the 46th International Symposium for Testing and Failure Analysis 2020, Pasadena, CA, USA, 15–19 November 2020; ASM International: Almere, The Netherlands, 2020; pp. 157–171.
5. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. Advances. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
6. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
7. Bao, H.; Dong, L.; Piao, S.; Wei, F. BEIT: BERT pre-training of image transformers. *arXiv* **2021**, arXiv:2106.08254.
8. Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; Sutskever, I. Generative pretraining from pixels. In Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 1691–1703.
9. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
10. Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; Hu, H. SimMIM: A simple framework for masked image modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9653–9663.
11. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
12. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Vitrtual, 13–18 July 2020; pp. 1597–1607.
13. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9912–9924.
14. Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.H.; Buchatskaya, E.; Doersch, C.; Pires, B.A.; Guo, Z.D.; Azar, M.G.; et al. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21271–21284.
15. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9650–9660.
16. Wu, Z.; Xiong, Y.; Yu, S.X.; Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3733–3742.
17. Ye, M.; Zhang, X.; Yuen, P.C.; Chang, S.F. Unsupervised embedding learning via invariant and spreading instance feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6210–6219.
18. Oord, A.V.D.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.

19. Tian, Y.; Krishnan, D.; Isola, P. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020, Proceedings, Part XI 16*; Springer International Publishing: Cham, Switzerland, 2020; pp. 776–794.
20. Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; Hinton, G.E. Big self-supervised models are strong semi-supervised learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 22243–22255.
21. Chen, X.; Fan, H.; Girshick, R.; He, K. Improved baselines with momentum contrastive learning. *arXiv* **2020**, arXiv:2003.04297.
22. Chen, X.; Xie, S.; He, K. An empirical study of training self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9640–9649.
23. Chen, X.; He, K. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 15750–15758.
24. Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 12310–12320.
25. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
26. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015, Proceedings, Part III 18*; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
27. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404. [CrossRef]
28. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
29. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
30. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3146–3154.
31. Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Lin, Z.; Liu, H. Expectation-maximization attention networks for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9167–9176.
32. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
33. Jain, J.; Singh, A.; Orlov, N.; Huang, Z.; Li, J.; Walton, S.; Shi, H. Semask: Semantically masked transformers for semantic segmentation. *arXiv* **2021**, arXiv:2112.12782.
34. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
35. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment anything. *arXiv* **2023**, arXiv:2304.02643.
36. Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; Huang, T. SegGPT: Segmenting everything in context. *arXiv* **2023**, arXiv:2304.03284.
37. Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Gao, J.; Lee, Y.J. Segment everything everywhere all at once. *arXiv* **2023**, arXiv:2304.06718.
38. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
39. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
40. Li, D.; Li, C.; Chen, C.; Zhao, Z. Semantic segmentation of a printed circuit board for component recognition based on depth images. *Sensors* **2020**, *20*, 5318. [CrossRef] [PubMed]
41. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
42. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434.