Memory Guided Denoising Student-Teacher Segmentation Model for Anomaly Detection

Abstract

Data collection in the industrial field is difficult and the types of anomalies cannot be predicted, model training is performed unsupervised. The main model in this paper is to perform unsupervised anomaly detection through feature-level knowledge distillation. Two models are proposed. One model is pre-trained on the ImageNet data set and has fixed weights as the Teacher model. The other model is used as the Student model without pre-training but to fit the intermediate features of the Teacher model during the training process.Through this process, the Teacher model is used only for the characteristic representation ability of the training data containing normal samples is distilled to the Student, but the Student has not learned the Teacher's ability to represent abnormalities. In the inference stage, it is judged whether the sample is abnormal based on the difference between the Teacher and the Student. However, in the process of model knowledge distillation, there are potential problems of insufficient distillation and teachers' weak restraint on students' learning. The student model have a phenomenon of "knowledge forgetting". Therefore, I introduced the memory module to enhance the degree of knowledge distillation and the CoordAtt module based on the coordinated attention mechanism to make students pay more attention to the correlation between pixels when denoising. For the segmentation network to uniformly process abnormal feature maps of different resolutions, I added the SE module based on the channel attention mechanism to pay more attention to the features that are more beneficial to model training. The above modifications improve the robustness of the model.

Keywords: Anomaly detection, knowledge distillation, Memory module, Attention mechanism.

1 Introduction

Anomaly detection is the basic task of detecting and locating abnormal samples.Anomaly detection in the industrial field is mainly used to detect abnormalities in images of industrial products, and improve production efficiency and productivity by detecting defective products during the detection process. Due to the increasing demand and applications in broad domains, such as risk management, compliance, security, financial surveillance, health and medical risk, and AI safety, anomaly detection plays increasingly important roles, highlighted in various communities including data mining, machine learning, computer vision and statistics [8].

At first, anomaly detection was often applied to tabular data, tending to detect abnormal data that was biased towards the overall data distribution. Later, it was applied to the field of computer vision, by adding noise to the CIFAR10 or MNIST data set to simulate abnormal data, and through convolutional neural network

model performs the task of anomaly detection.In 2019, Bergmann et al. proposed the MVTechAD dataset [1], which includes 15 types of objects involved in industrial production and conforms to real-world industrial production scenarios. Since then, the dataset has become a public open source dataset for anomaly detection models. Nowadays, the number of anomaly detection data sets is increasing, and the most widely used ones include MVTech3D, VisaAD, BTAD.Initially, the anomaly detection task only limited to image-level anomaly detection, but pixel-level anomaly detection or anomaly segmentation, is more suitable for the real production process. Therefore, models based on convolutional neural networks are more widely used.

Nowadays, anomaly detection mainly includes methods such as knowledge distillation and image reconstruction. The model involved this paper mainly achieves anomaly detection through knowledge distillation.The model transfers the learned normal sample features to the simple student model (weak generalization) through knowledge distillation through the complex teacher model (strong generalization) during training. The student model only learns the feature representation of the normal samples, and infers At this time, it is inferred whether the sample is abnormal based on the difference in characteristics learned by the teacher model and the student model on the abnormal sample.Image reconstruction is to add noise or mask occlusion to the training data, and then use an image reconstruction network to reconstruct the image. During training, the model only reconstructs normal images, but during inference, the model cannot reconstruct abnormal images. In the model inference stage, anomalies are detected and segmented based on the difference between the original image and the reconstructed image.

2 Related works

The initial applications of knowledge distillation were model compression and lightweight networks [5]. Deep neural networks have shown good performance in many tasks. However, as the performance increases, the scale of the model also increases, which is not conducive to deployment in practical applications. Therefore, we need to find a shallow network to simulate the output of the deep network to achieve similar performance.

Denoising Autoencoder (DAE) is a variant of the autoencoder specifically designed to learn useful features in data and denoise the input data. The basic idea is to force the model to learn a useful representation of the data by adding noise to the input data and then training the autoencoder to reconstruct clean input from the noisy data.The training process of denoising autoencoders forces the model to learn to model the noise in the input data while maintaining the ability to express useful features. This approach helps extract key features in the input data, and in some cases, denoising autoencoders are able to learn useful representations in unsupervised learning tasks.

2.1 Anomaly detection model based on knowledge distillation

Bergmann et.al applied the knowledge distillation framework to anomaly detection for the first time and proposed the Uniformed Students model [2].The network structure is shown in Figure 1 below.Through patch-level knowledge distillation, the image is mapped to the Embedding level through the convolutional network for cosine similarity loss calculation. However, this model increases the calculation amount and inference time of the model due to multiple Student models.

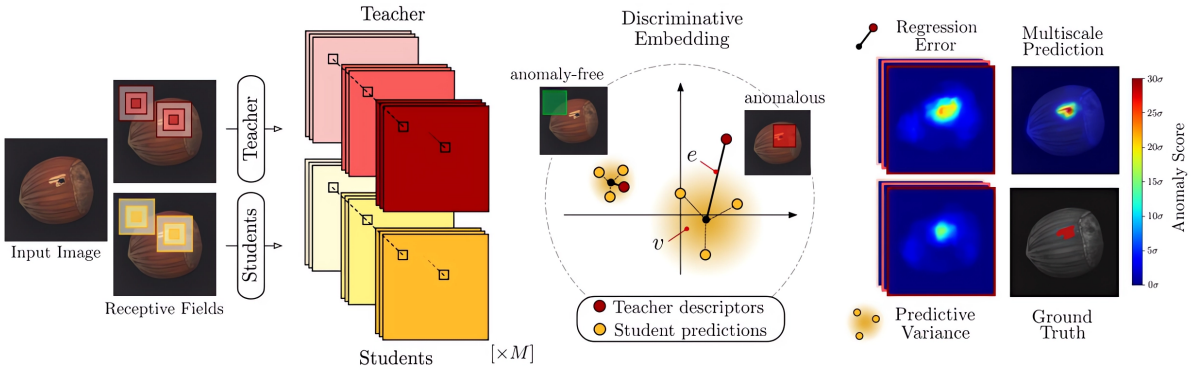

Figure 1. Uniformed Students model network structure

Subsequently, Salehi et.al proposed a feature-level knowledge distillation MKD model [10].This model enhances the granularity and degree of knowledge distillation by adding knowledge distillation to the feature layer between the Teacher model and the Student model. The model structure is shown in Figure 2 below. However, the problem that still exists with this model is that the Teacher model and the Student model have similar structures.

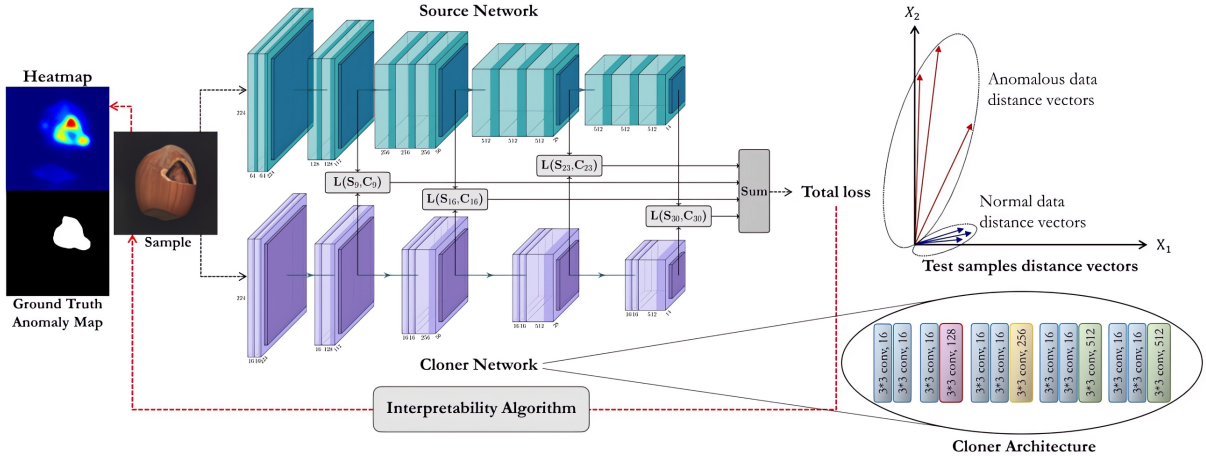

Figure 2. MKD model network structure

So later Hanqiu Denget.al proposed the reverse knowledge distillation RD model [3].This model performs knowledge distillation through a Teacher model with an encoder structure and a Student model with a decoder structure. The network structure is shown in Figure 3 below, which expands the difference between the two model structures and proposes a Bottleneck module, which can more It can effectively achieve multi-scale fusion of features and reduce feature redundancy.

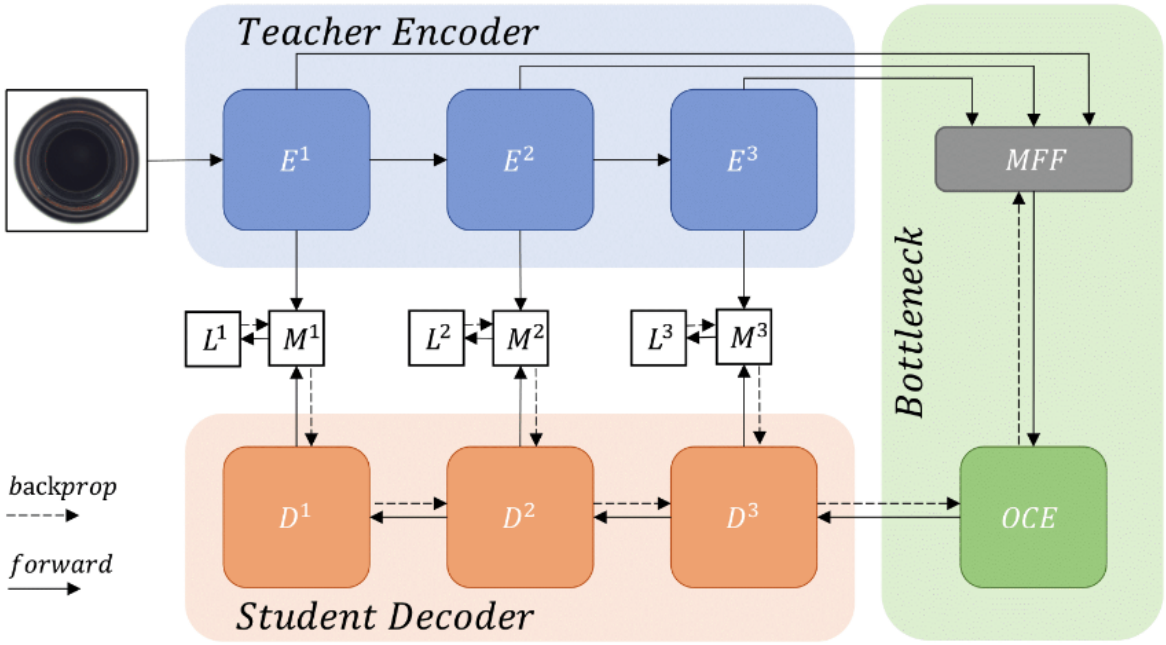

Figure 3. RD model network structure

The above methods all use unsupervised learning to train the network, but the accuracy problem needs to be solved. Therefore, some methods use noise as pseudo anomalies, thus adding semi-supervised learning training and improving the accuracy of the model.

2.2 Pseudo anomalies based on neural image reconstruction

In 2021, Zavrtanik et.al proposed the DRAEM [11]. This model is based on the image reconstruction method, but the accuracy of the model is improved by adding forged abnormal images. The model is mainly composed of two autoencoders. One is a reconstruction network, which is responsible for reconstructing the forged abnormal images back to normal images. The other model is a segmentation network similar to Unet [9]. This model reconstructs abnormal areas into normal areas. The network structure of the model is shown in Figure 4 below.

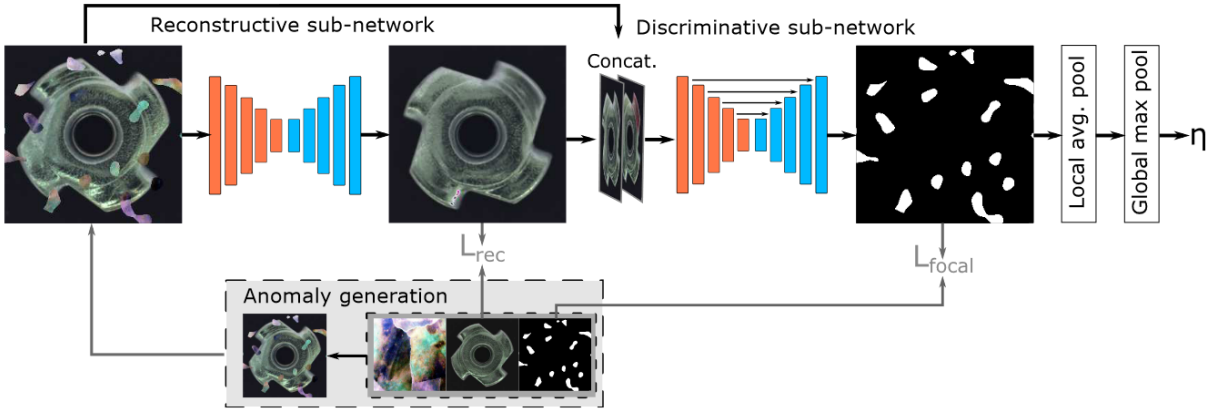

Figure 4. DRAEM network structure

After reconstruction, the abnormal image and the original abnormal image are sent to the segmentation network to segment the forged anomaly. The loss function of reconstruction training includes reconstruction loss MSE, segmentation loss SSIM and Focal loss. Because of the reconstruction of forged abnormal areas, the model's ability to represent normal samples is greatly enhanced, thereby improving the accuracy of model segmentation.

3 Method

The model DeSTSeg [12] selected in this paper summarizes the advantages of the two methods of knowledge distillation and image reconstruction. By adding a semi-supervised denoising process to the framework of knowledge distillation, the performance and accuracy of the model are improved.The denoising student teacher model adds Berlin noise forged abnormal images to the training process, so that the Student model has a "denoising" process for abnormal samples relative to the normal features output by the Teacher model, thus increasing the number of students and teachers in the inference stage. A better anomaly detection effect is achieved based on the differences in abnormal samples.

3.1 Overview

The overall DeSTSeg model consists of three sub-networks, namely Teacher network, Student network and segmentation network. The overall structure of the model is shown in Figure 5 below.

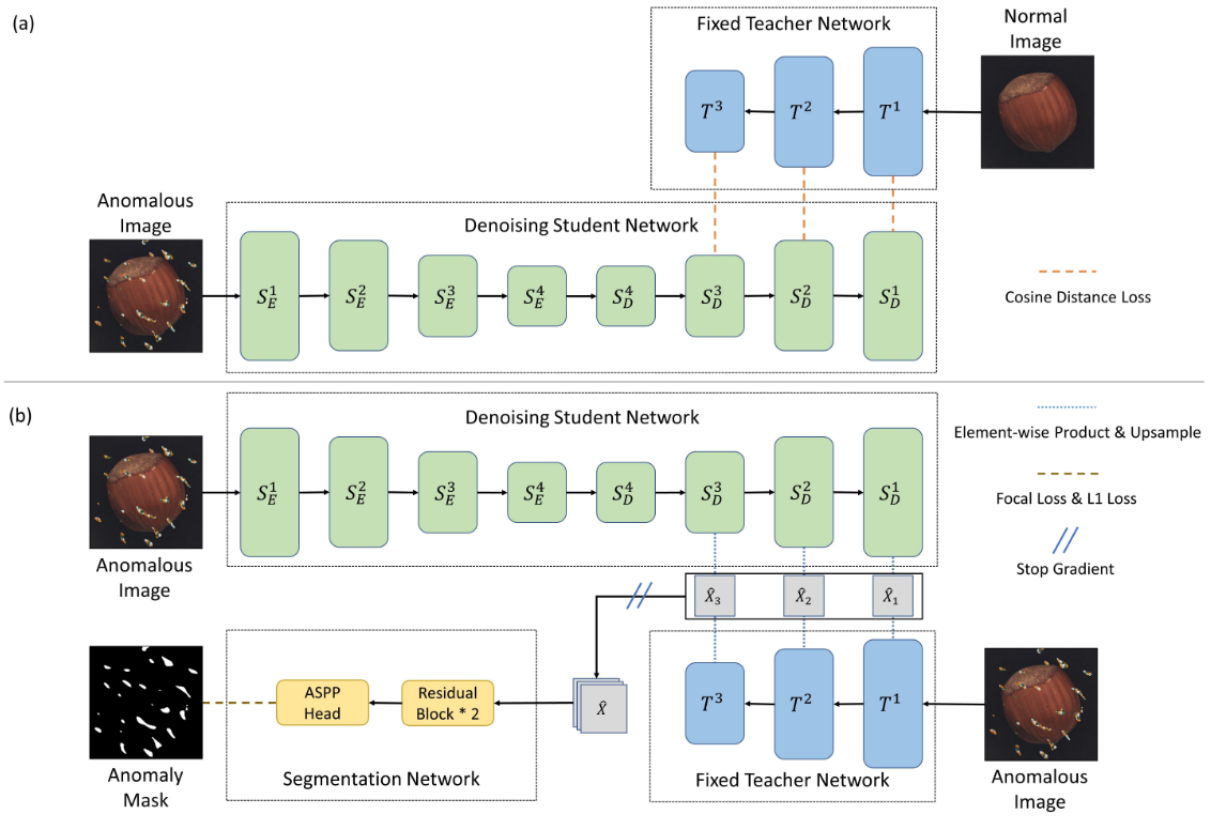

Figure 5. DeSTSeg model network structure

First, the student model and the teacher model conduct a reverse knowledge distillation training.The student model learns to align the features of the fake abnormal images to the features of the normal images of the teacher model.After training through knowledge distillation, the segmentation network starts training.The input of the segmentation network is the cosine similarity feature maps of the same image at different resolutions between the student model and the teacher model.The segmentation network segments these feature maps in areas close to 1 to obtain the final anomaly score feature map.

3.2 Pseudo anomalies generation

The abnormal image forged by Perlin noise passes through the Student model. Figure 5 is the process of generating pseudo anomalies, in which Perlin noise is used.Because the noise it generates is random in nature and produces continuous blocks, it is often used in the generation of game maps. Random noise P is generated through a Perlin noise generator.Then the anomaly mask $\mathrm{M_a}$ is generated through mean binarization.$\overline{\mathrm{M_a}}$ is obtained by inverting the anomaly mask.The overall generation process can be expressed as the following formula 1, where β is transparency. Under the influence of transparency parameters, abnormal generation is made more realistic.

$$\mathrm{I_a} = \overline{\mathrm{M_a}} \odot \mathrm{I} + (1-\beta)(\mathrm{M_a} \odot \mathrm{I}) + \beta(\mathrm{M_a} \odot \mathrm{A}) \tag{1}$$

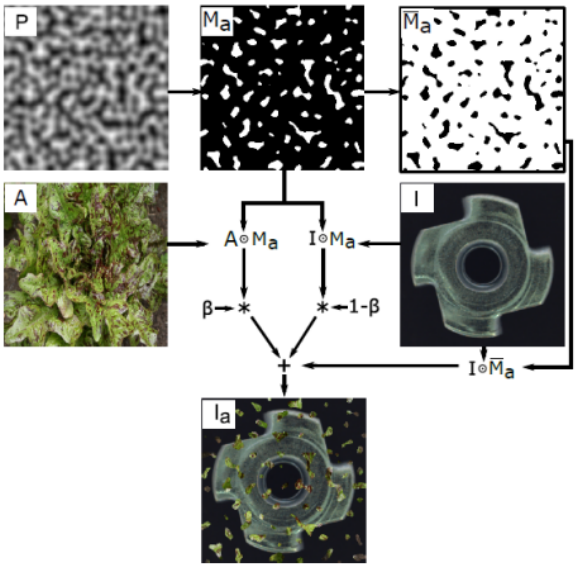

Figure 6. The process of pseudo anomalies generation

The abnormal source pictures are from the DTD dataset, while the non-abnormal pictures are from the MVTec dataset.

3.3 Denoise process based on knowledge distillation

In the decoder part of the Student model and the Teacher model, reverse knowledge distillation at the feature level is performed to calculate the cosine similarity loss. What the Student model wants to fit is the characteristics of normal images passing through the Teacher model. The model fits pseudo-abnormal features to normal features. The process is called "denoising".The denoising process is shown in Figure 7 below.

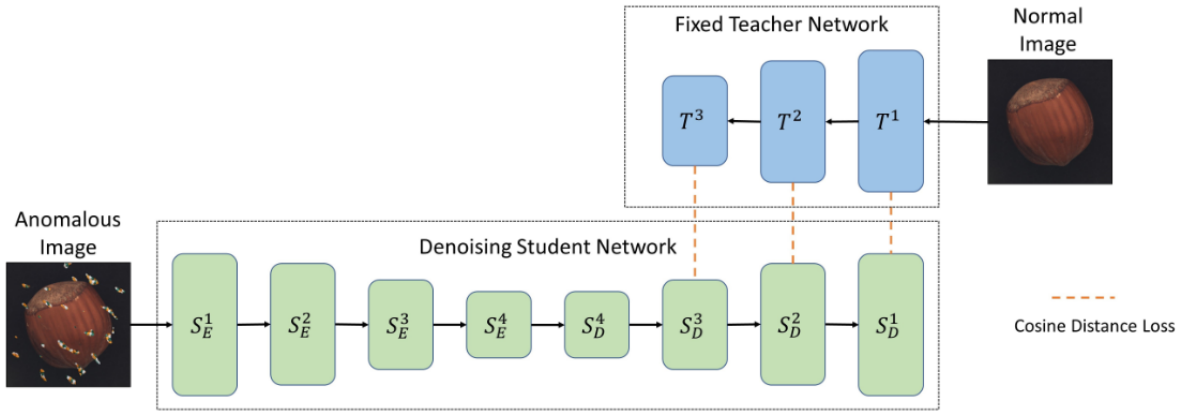

Figure 7. The process of Stduent model denoising

After the feature-level knowledge distillation converges, Student learns a representation that repairs the abnormal areas of the image during the encoding and decoding process.This process uses the calculation of cosine similarity loss, and the loss function formula is as follows:

$$X_k(i,j) = \frac{F_{T_k}(i,j) \odot F_{S_k}(i,j)}{\|F_{T_k}(i,j)\|_2 \|F_{S_k}(i,j)\|_2} \tag{2}$$

$$D_k(i,j) = 1 - \sum_{c=1}^{C_k} X_k(i,j)_c \tag{3}$$

$$L_{\cos} = \sum_{k=1}^{3} \left(\frac{1}{H_k W_k} \sum_{i,j=1}^{H_k,W_k} D_k(i,j) \right) \tag{4}$$

The $F_{T_k}, F_{S_k} \in R^{C_k \times H_k \times W_k}$ in the formula represents the feature of the middle layer between Student and Teacher. After using cosine similarity loss, the embedded feature output by the Student model moves closer to the Teacher.

3.4 Segmentation model training process

After the model converges in the denoising stage, the cosine similarity feature maps of the three-level features of the Teacher and Student decoders are upsampled to a unified size, and then spliced and passed into a simple segmentation network for training of the segmentation network.The segmentation network structure is very simple, consisting of two residual blocks and an ASPPHead. After the segmentation network, the anomaly score feature map of the sample is output, and the L1 loss and focal loss are calculated with the previously forged anomaly Mask. The process flow is shown in Figure 8 below.

During segmentation network training, the parameter updates of the previous Student model will be frozen, and only the parameters of the segmentation network will be updated.In the training set, most of the pixels are normal and easily recognized as background. Only a small portion of the image are pseudo-anomalous pixels that must be segmented. The model uses L1 loss and focal loss to make the segmentation network pay more attention to a small number of abnormal areas.The two loss calculation formulas are as follows:

$$L_{l1} = \frac{1}{H_1 W_1} \sum_{i,j=1}^{H_1,W_1} |M_{ij} - Y_{ij}| \tag{5}$$

$$p_{ij} = M_{ij} Y_{ij} + (1 - M_{ij})(1 - Y_{ij}) \tag{6}$$

$$L_{focal} = -\frac{1}{H_1 W_1} \sum_{i,j=1}^{H_1,W_1} (1 - p_{ij})^{\gamma} \log(p_{ij}) \tag{7}$$

$$L_{seg} = L_{focal} + L_{l1} \tag{8}$$

M is the anomaly score feature map output by the segmentation network and Y is the Mask label of the pseudo anomalies.

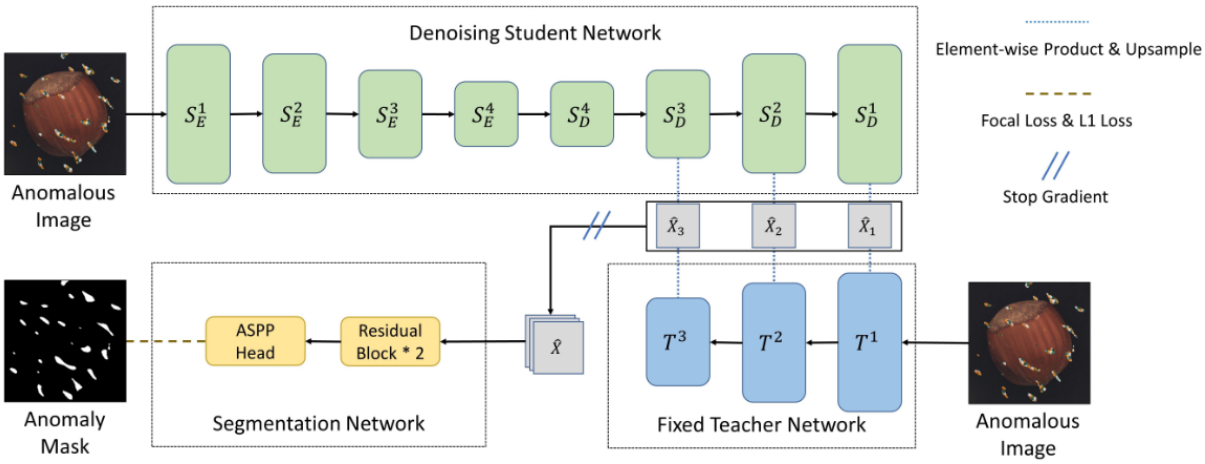

Figure 8. Segmentation network training process

The inference stage of the model is also the same as the segmentation network training process. The samples to be detected are passed into the student and teacher networks at the same time to obtain their pixel-level cosine similarity, and then passed into the segmentation network for abnormal segmentation.

4 Implementation details

When I ran the model code, I found that the cosine similarity loss in the knowledge distillation denoising stage had a “Abrupt Rise” phenomenon when training the segmentation network after stopping the gradient update.This shows that the model has not fully converged in the denoising stage, and the model has not thoroughly learned the ability to denoise the constantly changing pseudo-anomalies, which can also be called “knowledge forgetting”.

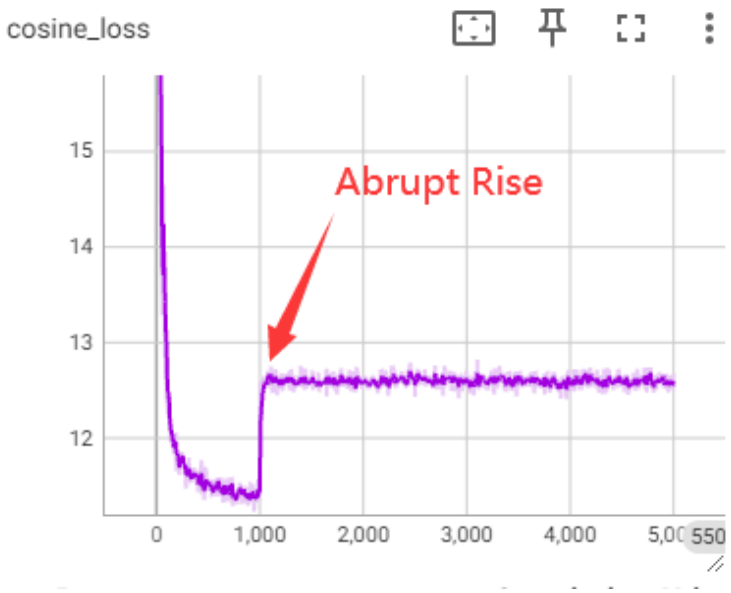

Figure 9. Insufficient distillation phenomenon

Moreover, the Teacher model only performs knowledge distillation on the decoder part of the student model, and does not have a good constraint on the encoder part.

4.1 Comparing with the released source codes

4.1.1 Improvement in knowledge distillation

Based on the above problems, a Memory Module was introduced, which originated from an anomaly detection work in 2019: MemAE [4].The MemAE model is a memory module created in the middle part of the autoencoder based on image reconstruction. The purpose is to memorize the characteristics of normal samples. Through the addressing and reorganization of features through the attention mechanism, the model can improve the representation of normal sample characteristics. The main structure is shown in Figure 10 below.

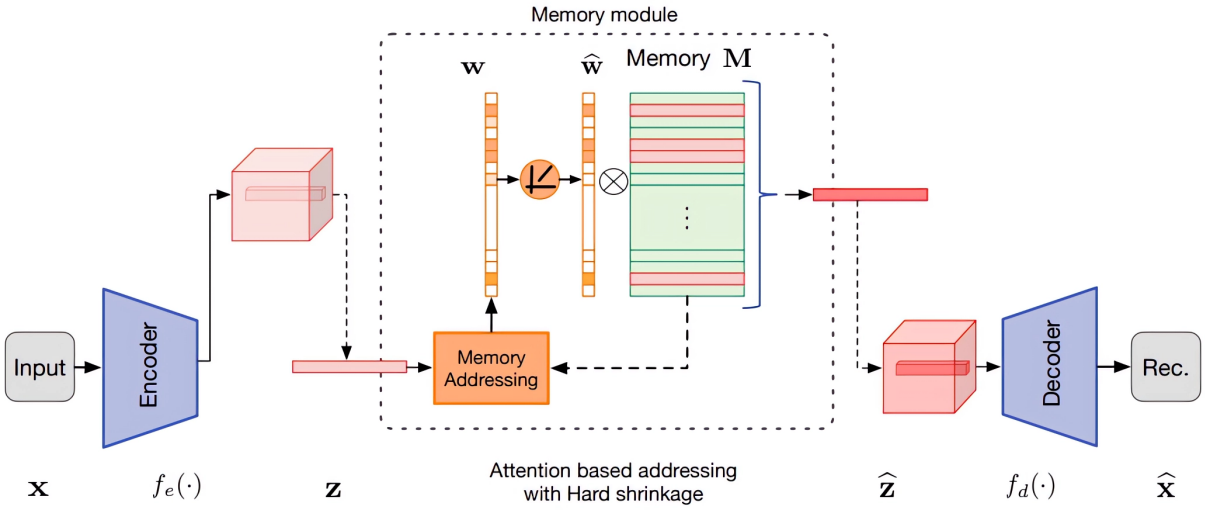

Figure 10. MemAE overall network structure

The Memory module stores a certain number of memory vectors $m_i \in \mathbb{R}^{1\times C}$.Through the attention mechanism, if there is a query$z \in \mathbb{R}^{C}$, calculate the cosine similarity between the query and each memory vector, and obtain the weight according to the similarity. Then, the weight is obtained according to the following formula, and the memory vectors are accumulated according to the weight to obtain the final value.

$$d(z, m_i) = \frac{z m_i^{\mathrm{T}}}{\|z\| \, \|m_i\|} \tag{9}$$

$$\omega_i = \frac{\exp(d(z, m_i))}{\sum_{j=1}^{N} \exp(d(z, m_j)} \tag{10}$$

$$\widehat{z} = \sum_{i=1}^{N} \omega_i m_i \tag{11}$$

After adding the Memory module to the Student middle layer and Teacher, through Teacher model training, an MSE loss calculation is performed on the reconstructed normal features and the original features. However, the experimental results are not good after doing so, and the performance of the previous model is reduced compared to the previous model. And it intensifies the "Abrupt Rise" of the model, as shown in Figure 10 below.

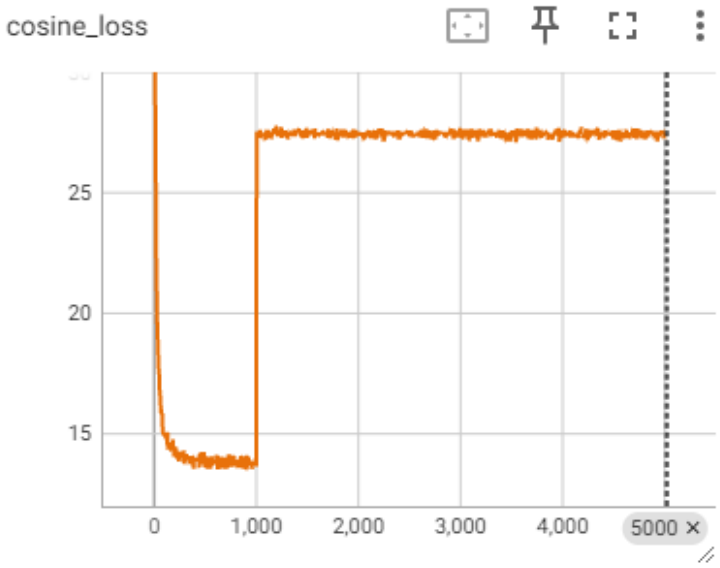

Figure 11. Add the Memory module directly without modification

I analyzed the reason and found that the original Memory paper used a reconstruction method, and the reconstruction loss MSE is not suitable for the knowledge distillation scheme of the DeSTSeg model.Therefore, Memory must be improved into a form suitable for knowledge distillation. After many experiments, it was found that the memory vectors in the Memory module should be reduced, the number of Memory modules should be increased and added to the middle of the network layer for feature extraction, and the MSE loss should be changed to cosine similarity loss.The overall network structure is shown in Figure 12 and Figure 13.

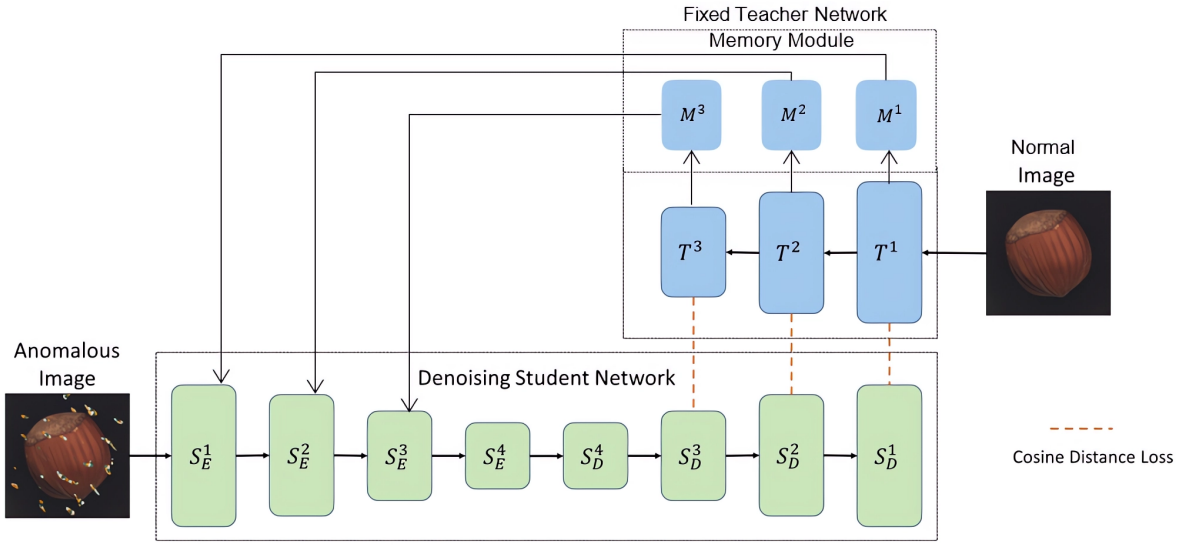

Figure 12. DeSTSeg model after adding Memory module

After the Teacher model is trained, the intermediate features output by the Student model are input into the Memory module for reorganization, thereby accelerating the convergence of the model and improving the degree of knowledge distillation.

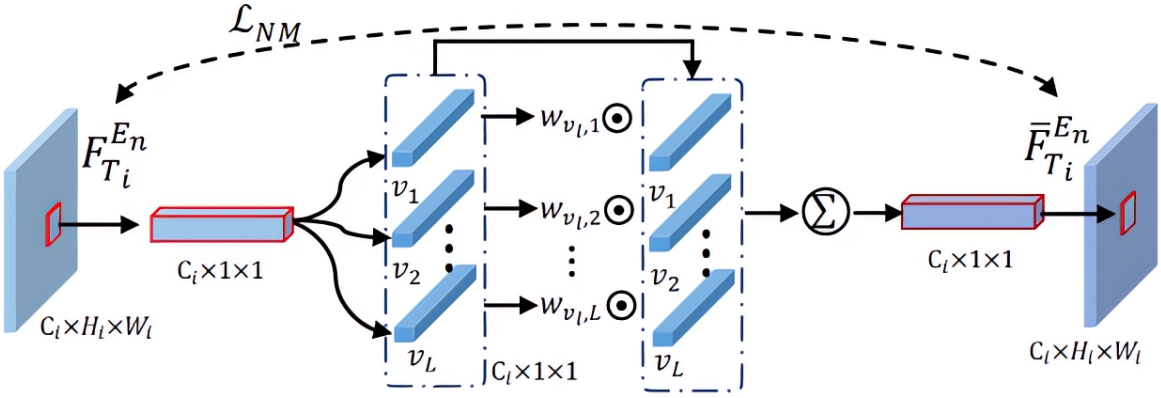

Figure 13. Memory module in the Teacher model training process

The data flow of the student model during the training phase will be recombined with different memory modules to enhance the degree of constraints of the teacher model on the student model encoder.

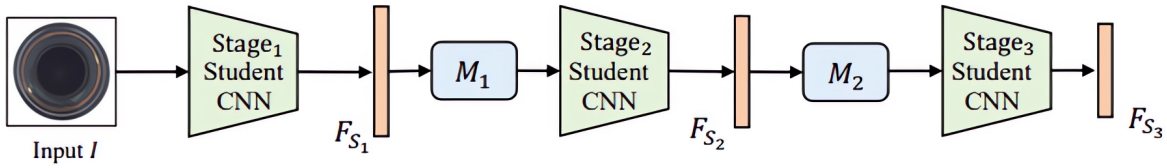

Figure 14. Memory module in the Teacher model training process

The loss function in the knowledge distillation stage is (λ is the hyperparameter of the balanced loss):

$$L_{KD} = L_{\cos} + \lambda L_{mem} \tag{12}$$

4.1.2 Improvement in Student module

In the data set, the sample images being detected are separated into backgrounds and objects. In the process of knowledge distillation, the features of the entire image in the network are regarded as an overall embedding. This approach ignores the positional information between pixels in the image.Therefore, the Coordinate attention module [6] (CA)is introduced to add an attention mechanism to the position information in features during knowledge distillation.

The CA module uses global pooling in the x and y directions to aggregate the input features in the vertical and horizontal directions into two independent direction-aware feature maps, and embeds the position information of the input Feature Map into the aggregated features of the channel attention vector.The two feature maps embedding direction-specific information are separately encoded into two attention maps, each attention map capturing the long-term dependence of the input feature map along one spatial direction. These two attention maps are then applied to the input feature map via multiplication to enhance the representation of the region of interest.The CA module structure is shown in Figure 15 below.

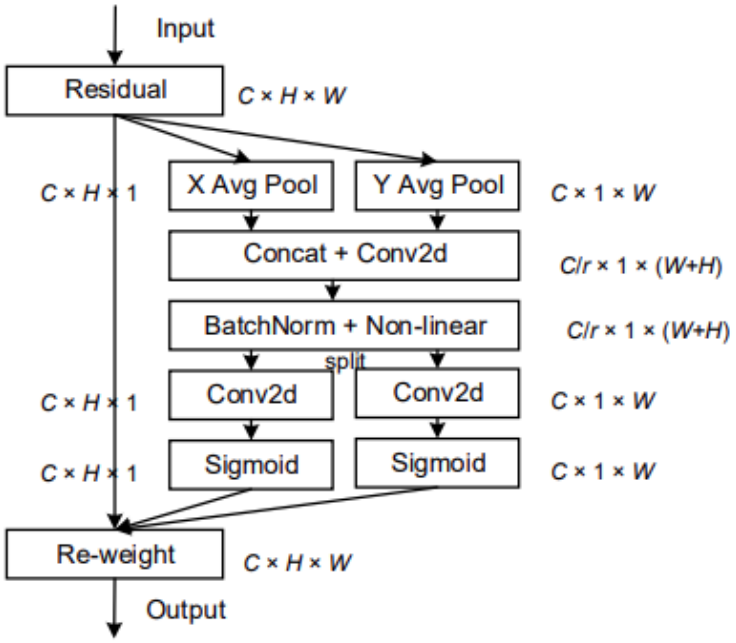

Figure 15. Coordinate attention module

Adding the CA module to the features to be distilled by the student model allows the student model to capture cross-channel information, as well as direction-awareness and position-awareness information, which can help the model more accurately locate and target objects of interest.

4.1.3 Improvement in Segmentation network

The input of the segmentation network is the cosine similarity feature map of the student network and the teacher network at three scales. These feature maps are spliced together in channel dimensions and become the input.Cosine feature maps of different scales contribute differently to the segmentation model. Feature maps with small sizes are more biased towards abstract differences at the semantic level, while feature maps with larger sizes are more biased towards low-dimensional differences at the texture level.The segmentation network should pay more attention to the scale information that is beneficial to the segmentation effect.Therefore, the SE module [7] based on the channel attention mechanism was introduced.The SE module structure is shown in the figure below.

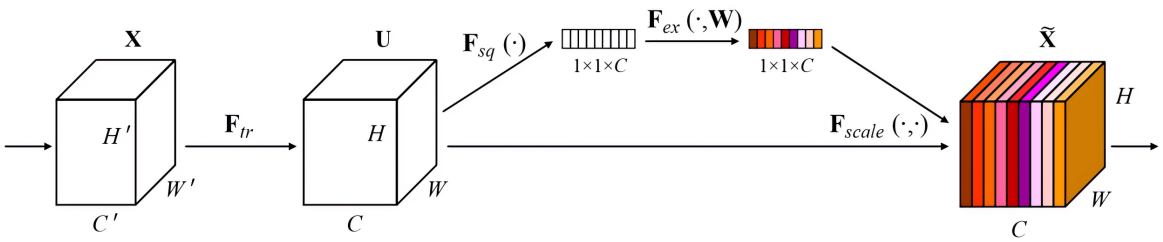

Figure 16. Coordinate attention module

By performing a global average pooling operation on the feature map, the SE module summarizes the information of the entire feature map into a feature vector. Doing so helps capture the relationships and interactions between different channels, allowing for a better modeling of the importance between channels.

4.2 Experimental environment setup

The experimental environment is an RTX3090 graphics card. The learning rate and optimizer adopt the settings of the original paper. The learning rate of the Memory module is set to 0.4, the hyperparameter λ of the Memory module loss function is set to 0.1, the knowledge distillation training is 1000 steps, and the segmentation network training is 4000 steps.

5 Results and analysis

First, after adding the above improved modules to the DeSTSeg model, experiments were conducted on the MVTechAD dataset. The experimental results are shown in Table 1 and Table 2 below.

Table 1. Image-level anomaly detection AUC

	DeSTSeg(origin)	DeSTSeg(memory)
bottle	99.2	**100**
cable	97.2	**97.7**
capsule	95.9	95.9
carpet	**99.2**	98.2
grid	100	100
hazelnut	**100**	99.2
leather	100	100
metal_nut	99.3	**99.6**
pill	**97.9**	97.5
screw	**92.7**	91.3
tile	**100**	99.2
toothbrush	99.7	99.7
transistor	96.1	**100**
wood	96.8	**99.1**
zipper	100	100
mean	98.3	**98.5**

Table 2. Pixel-level anomaly localization AUC

	DeSTSeg(origin)	DeSTSeg(memory)
bottle	99.0	**99.2**
cable	95.8	**97.3**
capsule	**98.9**	98.4
carpet	**98.2**	95.4
grid	**99.4**	98.7
hazelnut	99.6	99.6
leather	99.8	99.8
metal_nut	**98.9**	98.6
pill	**99.1**	98.9
screw	99.1	**99.3**
tile	**99.1**	98.0
toothbrush	**99.5**	99.4
transistor	90.1	**95.6**
wood	**96.3**	95.9
zipper	99.0	**99.2**
mean	98.1	**98.2**

It can be seen that the abnormal detection performance of transistor and wood objects in the data set has been significantly improved, and the abnormal segmentation ability of transistor and cable objects has been significantly improved. The overall improvement of the transistor object is the largest, and the protrusion phenomenon has also been improved.A certain degree of relief can be seen in Figure 17. Figure 18 below is the visualization of transistor anomaly detection.

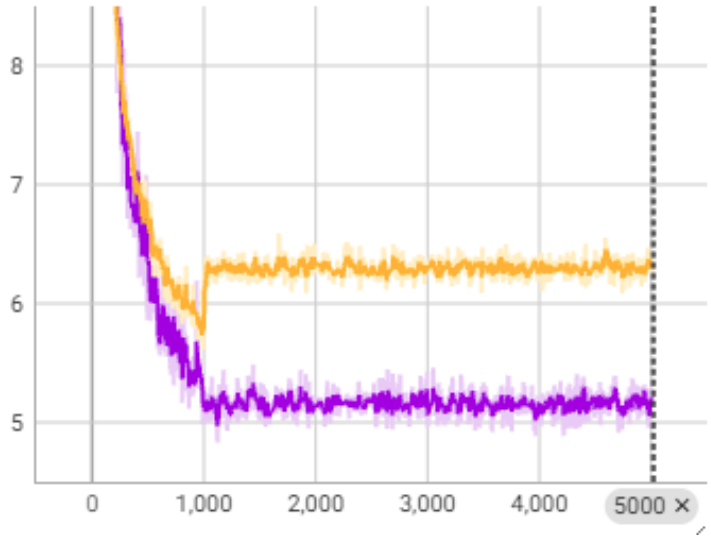

Figure 17. Purple is the knowledge distillation loss after adding the Memory module

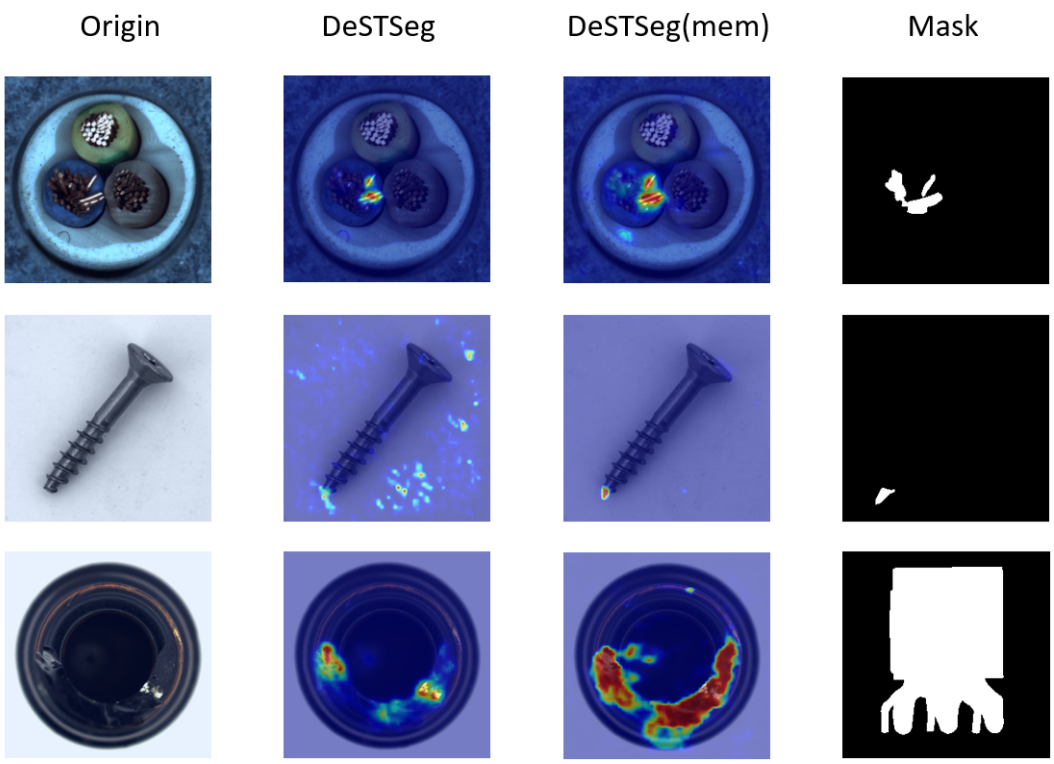

Figure 18. Samples abnormal segmentation visualization

Although the improved model has improved detection performance for certain types of objects, there are also some failure cases.It can be seen that how to use the information of normal samples still requires thinking and experimentation.

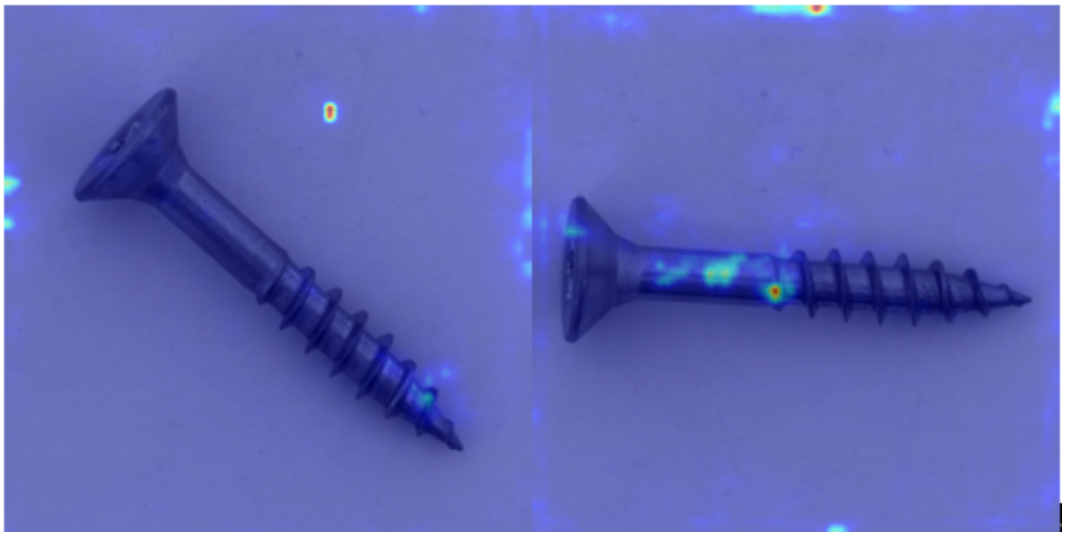

Figure 19. Samples abnormal segmentation visualization

The next improvements will also start from this aspect to improve the Student's ability to learn the subtle patterns of samples in order to solve the above-mentioned phenomenon.

6 Conclusion and future work

This article thinks about the problems encountered in the experiment of the model, and puts forward hypotheses based on the basis. It also looks at the problems encountered from multiple angles, fills in the potential shortcomings of the model, and improves the degree of knowledge distillation.And improvements were made based on the characteristics of the sub-networks of the model to improve the model's ability to utilize channel and spatial location information.However, the optimization of the model itself still needs to be thought about. Some details in the process have not yet been studied in depth. The final problems in the experimental part have yet to be solved. The next step will be to conduct further experiments and think about solutions, and solve the problem from the essence of the problem.

References

[1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020.

[3] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9737–9746, 2022.

[4] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019.

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[6] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722, 2021.

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[8] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[10] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912, 2021.

[11] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021.

[12] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3914–3923, 2023.