A Simplified Student Network with Multi-teacher Feature Fusion for Industrial Defect Detection

Mingjing Pei[1,2,3,4] and Ningzhong Liu[1,3,4](✉)

[1] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
lnz_nuaa@163.com
[2] School of Electronics and Information Engineering, West Anhui University, Lu'an, China
[3] MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing, China
[4] Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, China

Abstract. Improved industrial defect detection is deemed critical for ensuring high-quality manufacturing processes. Despite the effectiveness of knowledge distillation in detecting defects, there are still challenges in extracting useful features and designing a better student network. In this study, a new framework based on knowledge distillation is proposed to address these issues. To avoid overfitting and balance anomaly detection between image and pixel levels, a simplified student decoding network is designed. To extract more diverse features, a multi-teacher network is used in the teacher network. Simple addition and concatenation operations for feature maps were not sufficient, the method fuses features iteratively using attention mechanisms utilized for the multi-teacher networks. The approach is evaluated on two benchmark datasets for industrial defect detection, Mvtec and BTAD, and significantly improved performance is achieved compared to the other methods. An average accuracy of 99.22% and 97.46% for image-level and pixel-level ROC-AUC, respectively, is achieved on the Mvtec dataset, and 93.47% and 96.3% on the BTAD dataset. The proposed framework shows effectiveness in detecting defects in industrial settings, as demonstrated by the results.

Keywords: Industrial defect detection · Reverse knowledge distillation · Multi-teacher model

1 Introduction

Industrial defect detection is a technology that aims to identify defects in various industrial products to ensure their quality and maintain production stability. It has a wide range of applications, including unmanned quality inspection [1], intelligent inspection [2], and video surveillance [3]. However, industrial defect detection faces several challenges. One major challenge is related to data. Industrial defect detection requires large amounts of high-quality data to train and improve their accuracy. However, obtaining

H. Lu et al. (Eds.): ACPR 2023, LNCS 14406, pp. 245–258, 2023.
https://doi.org/10.1007/978-3-031-47634-1_19

such data can be challenging, as defects may be rare or unpredictable, and collecting and labeling data can be time-consuming and expensive. Another challenge is related to the complexity of industrial environments. Industrial defect detection must be able to identify defects in images with complex backgrounds, such as those with reflections or shadows, and in images with multiple types of defects. Finally, industrial defect detection must be highly accurate to detect subtle defects that may be difficult to identify with the human eye.

These challenges have led to the development of different approaches. It is usually divided into three categories. The first category is the synthesis of abnormal samples [4, 5]. These methods utilize neural networks for the binary classification of normal and synthetic abnormal samples. However, It requires pre-processing operations in the image, thereby increasing model complexity, and may not be able to effectively capture the complexity of real-world anomalies. The second category relies on pixel-level image comparison by reconstructing a normal image. Self-encoding and generative models, such as AE [6] and GAN [7], are commonly used in these methods. However, this approach may not align the reconstructed pixels with the input image and may alter the style of the image, leading to detection errors and performance degradation. The third category focuses on feature similarity, where the features extracted by convolutional neural networks, and anomalous regions have distinguishable feature embeddings [8]. These methods are more tolerant to noise interference and have better robustness in detection. Frameworks of teacher-student networks have been proposed on the basis of feature similarity [9], but they suffer from limitations such as insufficient feature extraction and the risk of overfitting due to using the same architecture.

To address these issues, a new network architecture has been designed. First, the insufficiency of extracting sufficiently diverse or representative features by using a single teacher network may be encountered. To tackle this problem, a multi-teacher network has been proposed to extract features with more diversity. Second, simple addition and concatenation operations may prove insufficient to achieve optimal performance when combining the learned features from multiple teacher networks. To overcome this problem, an iterative attention-based feature fusion method has been employed. Additionally, if the teacher and student network structures are identical, overfitting may occur. In such cases, the student network may overlearn to reconstruct abnormal regions, resulting in poor generalization performance during testing. To avoid this situation, a simplified student network has been designed with a different structure from the teacher network.

The key contributions of this work can be summarized as follows:

1. A network architecture consisting of multiple teachers and a student is utilized. The architecture extracts image features from multiple models.
2. Simple addition and concatenation operations for feature maps not be sufficient. To combine the features extracted from the multi-teacher networks, the iterative attention feature fusion method is employed. In order to prevent overfitting, the student and teacher networks use distinct architectures that are not consistent with each other.
3. To demonstrate the efficacy of our approach, numerous experiments were performed on industrial datasets.

2 Related Work

2.1 Reconstruction-Based Methods

Reconstruction-based methods are commonly used to detect anomalies by calculating pixel-level or image-level anomaly scores based on the reconstruction error, which is measured by models such as Auto-Encoders and Generative Adversarial Networks (GANs). However, these models can reconstruct abnormal samples well due to the generalization capability of neural networks, making it challenging to distinguish normal and abnormal samples accurately. To address this issue, researchers have modified the network architecture by incorporating memory mechanisms [10], generating pseudo-anomaly samples [11], and using image masking strategies [12]. These modifications aim to improve the ability to discriminate between normal and abnormal samples. However, current methods still have limited discrimination ability in real-world anomaly detection scenarios. To overcome these limitations, some researchers have proposed preprocessing techniques, such as Omni-frequency Channel-selection [13], to compress and restore individual images, and incorporate Position and Neighborhood Information to achieve better reconstruction results. While the reconstruction-based method is intuitive and has strong interpretability, existing methods can suffer from biases that result in detection errors and limit detection performance. For instance, the reconstructed image may have misaligned pixels or change the style of the original image. Additionally, the pixel-level comparison can be sensitive to noise, resulting in poor detection robustness.

2.2 Embedding-Based Methods

Embedding-based methods aim to detect anomalies by comparing the embedding vectors of normal and abnormal samples using a trained network. Commonly used approaches include defining embedding similarity, one-class classification [14, 16], and Gaussian distributions [15]. The literature [14] and [16] are popular methods that define a compact one-class distribution using normal sample vectors. To handle high-dimensional data, Deep SVDD has been adopted to estimate feature representations using deep networks. However, using the entire image embedding can only determine whether an image is normal or not, but cannot locate the abnormal regions. To address this issue, some researchers [17, 18] have used patch-level embeddings to produce an anomaly map, where each patch represents a point in the feature space. This approach allows for the localization of abnormal regions in the image. The literature [19] has proposed methods to obtain a better feature space representation by adapting the target dataset's features. Additionally, some researchers have used a decoupled hypersphere to further improve detection performance. Recently, a flow-based idea has been proposed to perform density estimation and calculate anomaly scores [20]. The goal of these methods is to find distinguishable embeddings, making them more robust to noise and other sources of interference.

2.3 Knowledge Distillation

Knowledge distillation networks typically consist of two networks. The student network imitates the output of the teacher network in order to simplify the model and reduce its resource requirements. In the field of anomaly detection, during the training phase, only normal samples are used, and through training, the feature vectors extracted by the teacher and student networks are very similar. During the testing phase, if an abnormal sample is encountered, the similarity of the feature vectors extracted by the two networks is very low. Using this idea, we can detect anomalies. Currently, based on this idea, there have been many works. For example, the framework proposed by the US method [21] includes a teacher network and multiple student networks. Through distillation of the last layer, multiple students imitate the output of the teacher network for anomaly detection. However, considering that the distillation of the last layer is insufficient, MKD [22] uses multiple feature layers for distillation, achieving good results. In order to further improve detection performance, reference [8] attempted to propose a reverse distillation architecture, compressing the extracted features through the teacher network and then decoding them through the student network, achieving good results. In addition, DeSTSeg [23] introduced the segmentation concept into the knowledge distillation architecture. This study is based on the reverse distillation framework to perform anomaly detection.

3 Method

The objective is to use a reverse distillation framework to detect anomalies in an input image by training a student network to imitate the behavior of multiple teachers' networks. By leveraging this approach, the student network can accurately identify anomalies in new images. In Fig. 1, in the training phase, firstly, there are three pre-trained teacher networks to extract the features of normal images, and then the features of different layers are fused by the multi-scale feature fusion module(MFF). Secondly, the output of the MFF module is fused using the iterative attention feature fusion module(iAFF), and then the obtained feature maps are compressed using the one-class embedding(OCE) module to obtain low-dimensional vectors. Finally, the low-dimensional vectors are decoded using the simplified student network to obtain the feature maps of each layer. The loss function is calculated by the feature layer corresponding to the teacher network. In the testing phase, the test images are modeled to obtain each feature layer of the simplified student network, and then the feature maps are fused to obtain the final anomaly results.

3.1 Multi-teacher Models

Most distillation network architectures are designed to distill knowledge from a one-teacher to one student. However, in the process of human learning, students often learn from multiple teachers. By incorporating multiple teacher models, a student model can benefit from multiple interpretations of the task. Ensemble learning with multiple teacher network predictions has been shown to outperform one teacher network. Therefore, we propose a distillation network architecture that utilizes multiple teachers to enhance the student model's performance.

Rich features are extracted from images using ResNet18, ResNet34, and Wide_ResNet50, which have been pre-trained on ImageNet. The model based on the resnet network family has a total of four convolutional blocks. We only used the previous

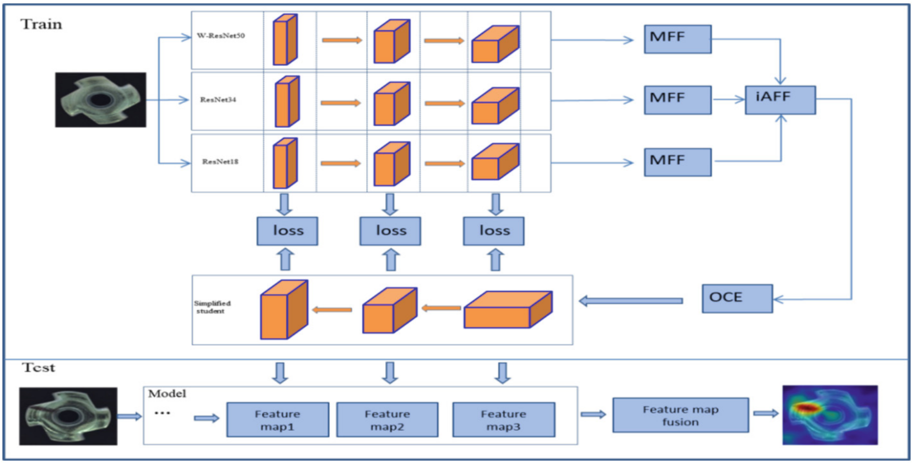

Fig. 1. An overview can be provided for the framework that has been proposed for the detection and localization of anomalies. The figure contains a training phase and a testing phase. In the training phase, there are three pre-trained teacher networks, feature fusion (MFF, iAFF), feature compression (OCE), and a simplified student network. In the testing phase, the feature maps of each feature layer of the student network are output by the model, and the visualization results can be obtained by the feature map fusion module.

three convolutional blocks to obtain the feature maps, removing the last convolutional block. To facilitate the use of a single decoding network in calculating the loss, the feature channels from the multiple teacher's networks need to be fused. This can be achieved through a simple concatenation operation. Once the feature channels have been fused, cosine similarity is used as the loss function for knowledge transfer.

During training, the input image is resized to 224 * 224 * 3. By fusion of the corresponding feature layers of the three networks, the first layer generates a feature map with dimensions of 64 * 64 * 384, the second layer generates a feature map with dimensions of 32 * 32 * 768, and the third layer generates a feature map of size 16 * 16 * 1536.

3.2 Feature Fusion and Dimensional Compression

Since each layer of each teacher network has different feature map sizes, feature map fusion needs to be done in order to get a uniform size. In addition, to effectively combine the learned features of a multi-teacher network, simple addition, and concatenation operations may not yield optimal results. As a solution, iterative attention feature fusion is adopted.

The MFF module is a type of module that facilitates multi-scale fusion by combining feature maps of varying sizes into a uniform size. This process results in a final feature map that is 16 * 16 in size and has a channel count of 3072. The IAFF module addresses short and long skip connections and the fusion of features caused within the Inception layer. Meanwhile, it is shown that the initial integration of feature maps can be a bottleneck and can be mitigated by adding another attention level. Specific details about IAFF can be found in the references [24]. After the iAFF module is applied, the feature map is resized to a size of 16 * 16 * 3072. The dimensionality of the feature vector is reduced by the OCE block, which employs ResNet's residual connection block. Further details about this module can be found in [9]. Ultimately, a feature map is 8 * 8 * 2048 is obtained as the output of the OCE block. Eventually, 3072 dimensions are compressed into 2048 dimensions.

3.3 Student Network

Some previous studies on knowledge distillation architectures for anomaly detection have employed identical architectures. However, this approach may lead to overfitting, as the student network can easily reconstruct the features of the anomaly region, thereby negatively impacting the detection performance. Therefore, a more suitable network architecture needs to be devised. Additionally, to ensure compatibility between the feature channels of the intermediate layers in the multiple teacher networks, a simplified decoding network is designed. The structure is illustrated in Fig. 2.

Figure 2 is a simplified student decoding network. The right side of the figure shows the student decoding network and the left side shows the basic blocks. There are a total of three layers in the student decoding network, and each layer contains the basic blocks.

The Residual Connection Block from ResNet [25] serves as the inspiration for the basic blocks, which use to combine feature maps of different scales. The first branch includes a deconvolution layer and a BatchNorm layer, it can keep original features. The second branch consists of two 1 * 1 convolutions and a 3 * 3 convolution. In this procedure, the 1 * 1 operation is first performed on the image, then the 1 * 1 convolution and 3 * 3 convolution are passed respectively, and finally, the concat operation is performed. The purpose is to reduce the amount of computation. To match the feature map size of the teacher network, a 4 * 4 deconvolution layer is applied in one branch of the Basic Block. The output of the two branches is combined through summation and then passed through a ReLU layer.

To match the structure of the teacher network, the student decoder module has the same three-layer structure, but in an inverted order. The D3 module is constructed with a Basic Block and a convolutional layer. The 1 * 1 convolutional layer is responsible for adjusting the number of input channels to match that of the corresponding teacher network's output channels. On the other hand, the D2 and D1 modules both consist of a Basic Block, which is used to maintain the original features and fuse feature maps of different scales.

3.4 Loss Function

Minimizing this loss function is the main objective of the training process. In out model, the formula for calculating a two-dimensional anomaly map [27] is as follows.

$$M^k(h, w) = 1 - \frac{f_T^k(h, w) f_S^k(h, w)}{\left\| f_T^k(h, w) \right\| \cdot \left\| f_S^k(h, w) \right\|} \tag{1}$$

The height and width of the feature map are denoted by h and w, respectively, while the number of feature layers is represented by k. The variables $f_T^k(h, w)$ and $f_S^k(h, w)$

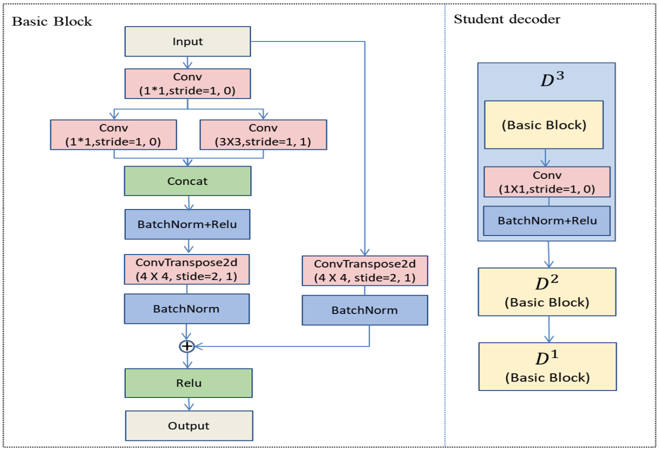

Fig. 2. A simplified student decoding network.

denote the feature vectors. The loss function is defined as shown in (2).

$$L = \sum_{k=1}^{3} \left\{ \frac{1}{H_k W_k} \sum_{h=1}^{H_k} \sum_{w=1}^{W_k} M^k(h, w) \right\} \tag{2}$$

4 Experiments

4.1 Datasets

The MVTec AD dataset [27] has been developed with a focus on industrial inspection, and it serves as a means of comparing different methods for anomaly detection. The dataset comprises 15 categories of objects and textures, with each class being further divided into a training set and a testing set. The training set consists of 3629 images of typical objects, while the test set includes 1725 images.

As for the BTAD dataset [28], it contains 2830 images showcasing actual industrial anomalies in 3 different products. The dataset contains 3 categories, including 2 for objects and 1 for texture, with 1800 images for training and 741 images for testing.

4.2 Implementation Details

The experiments were carried out on an Ubuntu 18.04 operating system using an NVIDIA GTX Geforce 2070s GPU. In these experiments, we employed pre-trained ResNet18, ResNet34, and Wide_ResNet50 as the backbone. All images in the MVTec and BTAD datasets were resized to 256 * 256 pixels. To optimize our model, we used the Adam optimizer. The learning rate was scheduled using the cosine annealing strategy. We trained our model for 200 epochs, using a batch size of 4.

4.3 Results

To assess the effectiveness of our model for image-level anomaly detection, the performance was compared with existing methods. Table 1 presents the results of our image-level anomaly detection, indicating that our model achieved superior performance compared to other models, with a score of 99.22. Moreover, our method also demonstrated better results in terms of texture and object class, reaching 99.54 and 99.9, respectively, when compared to other models. These findings suggest that our model is an effective approach to image anomaly detection.

Furthermore, Our model showed remarkable improvement compared to other models in performance, indicating the positive impact of our multi-teacher networks, feature fusion, and a simplified student decoding network. Our method differs from the US method in that we utilize an inverse teacher-student network architecture, while the US does not. We hypothesize that in the absence of an inverse structure, the student network may not effectively identify anomalous regions. Notably, some categories such as Grid, Pill, and Zipper demonstrated better results than our method, although our performance was comparable to theirs.

For pixel-level anomaly detection, we compared our method with several baselines including SPADE, Cutpaste, Draem, UniAD, NSA, and ADRD. However, it should be noted that our comparison did not encompass all the methods presented in the image-level anomaly detection evaluation, and there were instances where some data were not disclosed in the publication.

Based on the results presented in Table 2, it is evident that our method achieved a good performance in pixel-level anomaly detection, with a ROC-AUC score of 97.46, indicating its effectiveness. Our method achieves optimal results in some categories, e.g. Transistor, and Zipper. The Draem method achieves good results in some categories where it is based on pixel-level segmentation. This provides a good idea for pixel-level localization.

Table 1. Results of image-level anomaly detection on the MVTec dataset.

Category		RIAD	US	PaDiM	Cutpaste	NSA	Draem	UniAD	ADRD	Ours
Texture	Carpet	84.2	91.6	99.8	93.9	95.6	97	99.9	98.9	**100**
	Grid	99.6	81	96.7	**100**	99.9	99.9	98.5	**100**	99.8
	Leather	100	88.2	**100**	**100**	99.9	**100**	**100**	**100**	**100**
	Tile	98.7	99.1	98.1	94.6	100	**99.6**	99	99.3	98.2
	Wood	93	97.7	99.2	99.1	97.5	99.1	97.9	99.2	99.7
	Average	95.1	91.5	98.8	97.5	98.6	99.12	99.06	99.48	**99.54**
Object	Bottle	99.9	99.0	99.9	98.2	97.7	99.2	**100**	**100**	**100**
	Cable	81.9	86.2	92.7	81.2	94.5	91.8	97.6	95	**98.6**

(*continued*)

Table 1. (*continued*)

Category		RIAD	US	PaDiM	Cutpaste	NSA	Draem	UniAD	ADRD	Ours
	Capsule	88.4	86.1	91.3	98.2	95.2	**98.5**	85.3	96.3	97.6
	Hazelnut	83.3	93.1	92	98.3	94.7	**100**	99.9	99.9	**100**
	Metal nut	88.5	82.0	98.7	99.9	98.7	98.7	99	**100**	**100**
	Pill	83.8	87.9	93.3	94.9	99.2	**98.9**	88.3	96.6	97
	Screw	84.5	54.9	85.8	88.7	90.2	93.9	91.9	97	**97.8**
	Toothbrush	100	95.3	96.1	99.4	100	**100**	95	99.5	**100**
	Transistor	90.9	81.8	97.4	96.1	95.1	93.1	**100**	96.7	98.3
	Zipper	98.1	91.9	90.3	99.9	99.8	**100**	96.7	98.5	99.7
	Average	89.9	85.8	93.8	95.5	96.5	97.41	95.37	97.95	**98.9**
Average		91.7	87.7	95.5	96.1	97.2	98.27	96.6	98.72	**99.22**

The VT-ADL and autoencoder method were also compared with our model on the BTAD dataset. It can be observed from Table 3 that other models were outperformed by our method. Since pixel-level anomaly detection scores were not provided by other methods, they could not be compared. In categories 1 and 2, better results than our method were achieved by the other method.

Figure 3 shows the visualization results of our models. The red area indicates the area considered by the model to be the higher anomaly region. We show the original image, the mask image, and the visualization result separately. From the results, we can see that our model can accurately detect the defective regions, which indicates that our model has good detection performance.

Table 2. ROC-AUC results for pixel-level anomaly detection on the MVTec AD dataset are present-ed. The best-performing method for each category is highlighted in bold.

Category		SPADE	Cutpaste	Draem	NSA	UniAD	ADRD	Ours
Texture	Carpet	97.5	98.3	96.2	95.5	98	98.9	**99.3**
	Grid	93.7	97.5	**99.5**	99.2	94.6	99.3	99.2
	Leather	97.6	99.5	98.9	99.5	98.3	99.4	99.4
	Tile	87.4	90.5	**99.5**	99.3	91.8	95.6	94.5
	Wood	88.5	95.5	**97**	90.7	93.4	95.3	94
	Average	92.9	96.3	**98.2**	96.8	97.8	97.7	97.28
Object	Bottle	98.4	97.6	**99.3**	98.3	98.1	98.7	98.4
	Cable	97.2	90	95.4	96	96.8	**97.4**	97.3

(*continued*)

Table 2. (*continued*)

Category		SPADE	Cutpaste	Draem	NSA	UniAD	ADRD	Ours
	Capsule	99.0	97.4	94.1	97.6	97.9	98.7	98.4
	Hazelnut	99.1	97.3	**99.5**	97.6	98.8	98.9	99
	Metal nut	98.1	93.1	**98.7**	98.4	95.7	97.3	96.6
	Pill	96.5	95.7	97.6	98.5	95.1	98.2	96.1
	Screw	98.9	96.7	**99.7**	96.5	97.4	99.6	99.3
	Toothbrush	97.9	98.1	98.1	94.9	97.8	**99.1**	99
	Transistor	94.1	93.0	90.0	88	98.7	92.5	**93.5**
	Zipper	96.5	99.3	98.6	94.2	96.0	98.2	**98.7**
	Average	97.6	95.7	97.1	96	97	**97.86**	97.63
Average		96.5	96	97.5	96.3	97.4	**97.78**	97.46

Table 3. Results of image-level and pixel-level anomaly detection for the BTAD dataset. The symbol "-" indicates that the method does not provide a pixel-level anomaly score.

BTAD	AE + MSE	AE + MSE + SSIM	VT-ADL	Ours
0	49/-	53/-	99/-	**100**/99.6
1	92/-	**96**/-	94/-	94.6/92.3
2	95/-	**89**/-	77/-	85.8/97
Mean	78/-	79/-	90/-	**93.47**/96.3

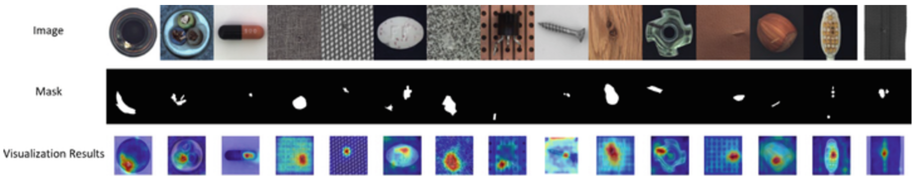

Fig. 3. Visualization results of image anomaly detection.

In Fig. 4, statistics of the abnormal scores were presented. The results indicate that the model could effectively distinguish between normal and abnormal samples. However, there were some categories where the distinction was not clear, indicating the need for an improved model.

4.4 Ablation Study

Table 4 shows the impact of pretraining and non-pretraining networks on image-level and pixel-level anomaly detection results, demonstrating the significance of pretraining

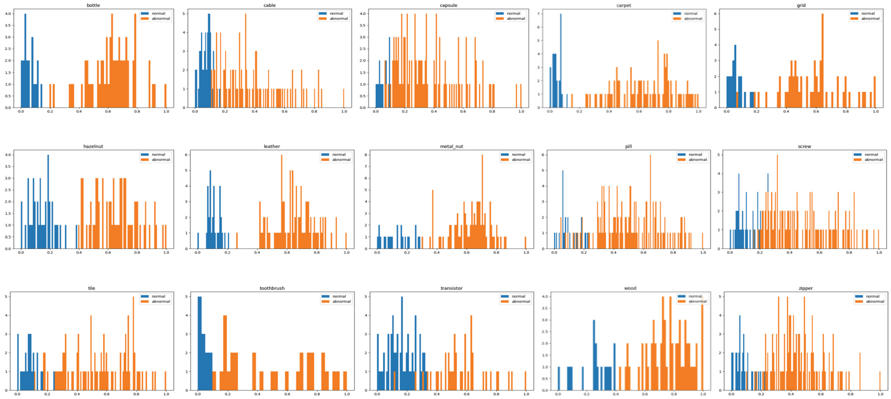

Fig. 4. The abnormal scores were statistically analyzed, and the distinction between normal and abnormal samples was displayed in the histogram. Normal samples were represented by blue bars, while abnormal samples were represented by yellow bars.

in image feature extraction. Results reveal that utilizing pretraining can enhance the accuracy by approximately 25 points, suggesting that pretraining is effective in capturing image features.

Table 4. Results of anomaly detection using our model, comparing the performance of models trained with and without pretraining.

Three Teachers Network	Unpretrained	Pretrained
Texture	77.42/70.12	99.7/97.28
Object	68.94/78.17	98.9/97.63
Average	73.18/74.15	99.22/97.46

Table 5 presents the results of different modifications made to the original network for image-level and pixel-level anomaly detection. The baseline represents the original network. The simplified student network corresponds to our own designed student network, which replaces the original one. The multi-teachers network involves using three teacher networks instead of one. For the feature fusion of multiple teachers, the attention iteration module was introduced.

Employing different architectures for the teacher and student networks can lead to improved results. This highlights the potential risk of overfitting when using a high-performing student network, which may inadvertently recover anomalies.

The adoption of multi-teacher networks can enhance the extraction of feature representations, as evidenced in Table 5, where the image-level anomaly detection results improve significantly, while the pixel-level anomaly detection results remain relatively unchanged.

Table 5. Evaluating the effectiveness of the proposed networks and modules in detecting anomalies.

Baseline	Simplified Student	Three Teachers	iAFF	Value
√				98.72/97.78
√	√			98.77/97.93
√	√	√		98.91/97.56
√	√	√	√	99.22/97.46

Adopting iterative attention feature fusion can further enhance the quality of feature extraction. Table 5 demonstrates that the employment of these modules leads to improved results in image-level anomaly detection tasks.

5 Conclusion

A new framework for industrial defect detection is proposed, which involves the use of reverse knowledge distillation with three teachers. To prevent overfitting and balance image and pixel detection, a simplified student decoding network was created. To extract more features, a multi-teacher network was used for the teacher network. However, simple addition and concatenation of feature maps were found to be inadequate, and therefore an iterative attention feature fusion method was employed to effectively combine the features of the multi-teacher networks. Extensive experiments are conducted on two datasets, Mvtec and BTAD, and it is demonstrated that the anomaly detection scores of image-level tasks are improved by our approach. However, the improvement in pixel-level performance has decreased, and further investigation is needed to enhance it in the future.

Acknowledgements. This research is supported in part by the Natural Science Foundation of Jiangsu Province of China (BK20222012), Guangxi Science and Technology Project (AB22080026/2021AB22167), National Natural Science Foundation of China (No. 61375021) and the Natural Science Key Project of Anhui Provincial Education Department (No. KJ2020A0636, No. KJ2021A0937, No. 2022AH051683, No. 2022AH051670, No. 2022AH051669).

References

1. Lee, J.H., et al.: A new image-quality evaluating and enhancing methodology for bridge inspection using an unmanned aerial vehicle. Smart Struct. Syst. **27**(2), 209–226 (2021)
2. Ullah, W., et al.: Artificial intelligence of things-assisted two-stream neural network for anomaly detection in surveillance big video data. Futur. Gener. Comput. Syst.. Gener. Comput. Syst. **129**, 286–297 (2022)
3. Patrikar, D.R., Parate, M.R.: Anomaly detection using edge computing in video surveillance system. Int. J. Multimedia Inf. Retrieval **11**(2), 85–110 (2022)

4. Li, C.-L., Sohn, K., Yoon, J., Pfister, T.: Cutpaste: self-supervised learning for anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9664–9674 (2021)
5. Schlüter, H.M., Tan, J., Hou, B., Kainz, B.: Natural synthetic anomalies for self-supervised anomaly detection and localization. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13691, pp. 474–489. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19821-2_27
6. Wang, S., et al.: Auto-AD: autonomous hyperspectral anomaly detection network based on fully convolutional autoencoder. IEEE Trans. Geosci. Remote Sens.Geosci. Remote Sens. **60**, 1–14 (2021)
7. Lu, H., Du, M., Qian, K., He, X., Wang, K.: GAN-based data augmentation strategy for sensor anomaly detection in industrial robots. IEEE Sens. J. **22**(18), 17464–17474 (2021)
8. Pei, M., Liu, N., Gao, P., Sun, H.: Reverse knowledge distillation with two teachers for industrial defect detection. Appl. Sci. **13**(6), 3838 (2023)
9. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9737–9746 (2022)
10. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14318–14328 (2022)
11. Ding, C., Pang, G., Shen, C.: Catching both gray and black swans: open-set supervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7388–7398 (2022)
12. Ristea, N.-C., et al.: Self-supervised predictive convolutional attentive block for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13576–13586 (2022)
13. Liang, Y., Zhang, J., Zhao, S., Wu, R., Liu, Y., Pan, S.: Omni-frequency channel-selection representations for unsupervised anomaly detection. arXiv preprint arXiv:2203.00259 (2022)
14. Chen, Y., Tian, Y., Pang, G., Carneiro, G.: Deep one-class classification via interpolated gaussian descriptor. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 1, pp. 383–392 (2022)
15. Sun, X., Yang, Z., Zhang, C., Ling, K.-V., Peng, G.: Conditional Gaussian distribution learning for open set recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13480–13489 (2020)
16. Zhang, F., Fan, H., Wang, R., Li, Z., Liang, T.: Deep dual support vector data description for anomaly detection on attributed networks. Int. J. Intell. Syst. **37**(2), 1509–1528 (2022)
17. Defard, T., Setkov, A., Loesch, A., Audigier, R.: Padim: a patch distribution modeling framework for anomaly detection and localization. In: Del Bimbo, A., et al. (eds.) ICPR 2021. LNCS, vol. 12664, pp. 475–489. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-68799-1_35
18. Wang, S., Wu, L., Cui, L., Shen, Y.: Glancing at the patch: anomaly localization with global and local feature comparison. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 254–263 (2022)
19. Lee, S., Lee, S., Song, B.C.: CFA: coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. IEEE Access **10**, 78446–78454 (2022)
20. Yu, J., et al.: Fastflow: unsupervised anomaly detection and localization via 2D normalizing flows. arXiv preprint arXiv:2111.07677 (2021)
21. Wang, G., Han, S., Ding, E., Huang, D.: Student-teacher feature pyramid matching for anomaly detection. arXiv preprint arXiv:2103.04257 (2021)

22. Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M.H., Rabiee, H.R.: Multiresolution knowledge distillation for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14902–14912 (2021)
23. Zhang, X., Li, S., Li, X., Huang, P., Shan, J., Chen, T.: DeSTSeg: segmentation guided denoising student-teacher for anomaly detection. arXiv preprint arXiv:2211.11317 (2022)
24. Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., Barnard, K.: Attentional feature fusion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3560–3569 (2021)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
26. Yu, J., Liu, J.: Two-dimensional principal component analysis-based convolutional autoencoder for wafer map defect detection. IEEE Trans. Ind. Electron. **68**(9), 8789–8797 (2020)
27. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: MVTec AD–a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9592–9600 (2019)
28. Mishra, P., Verk, R., Fornasier, D., Piciarelli, C., Foresti, G.L.: VT-ADL: a vision transformer network for image anomaly detection and localization. In: 2021 IEEE 30th International Symposium on Industrial Electronics (ISIE), pp. 01–06. IEEE (2021)