Multimedia Tools and Applications (2024) 83:90573–90594
https://doi.org/10.1007/s11042-024-18512-5

1247: RECENT ADVANCES IN AI-POWERED MULTIMEDIA VISUAL COMPUTING AND MULTIMODAL SIGNAL PROCESSING FOR METAVERSE ERA

Masked feature regeneration based asymmetric student–teacher network for anomaly detection

Haocheng Gu[1] · Gongyang Li[2] · Zhi Liu[1]

Received: 23 October 2023 / Revised: 22 December 2023 / Accepted: 29 January 2024 /
Published online: 15 February 2024

Abstract

Anomaly detection (AD), also known as one-class classification and localization, has become a challenging task in the field of industrial defect detection where anomalous samples can hardly be collected to train a network based on conventional computer vision tasks. Knowledge distillation based on a student-teacher (S-T) framework has proved its effectiveness in solving such problems in an unsupervised fashion. In this paper, we propose a novel asymmetric S-T framework with masked feature regeneration called AST-MFR. First, to ensure better feature alignment during the training period, we introduce a masked feature regeneration (MFR) module to mask multi-level features of the student network randomly and regenerate the corresponding features under the guidance of the teacher network's features. Second, to enlarge the feature diversity of unseen anomalous samples during the test period, we adopt an asymmetric S-T network structure that is sensitive to detecting and locating anomalous parts. We conduct experiments on the industrial anomaly detection benchmark dataset MVTec AD and the results demonstrate the proposed model achieves competitive performance compared to the state-of-the-art methods on both anomaly detection and anomaly localization .

Keywords Anomaly detection · Knowledge distillation · Unsupervised learning · Masked feature regeneration · Asymmetric student-teacher framework

✉ Zhi Liu
liuzhisjtu@163.com

Haocheng Gu
guhaocheng0528@163.com

Gongyang Li
ligongyang@shu.edu.cn

[1] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China

[2] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China

1 Introduction

Metaverse technology is a comprehensive application based on various technologies such as artificial intelligence, virtual reality, and augmented reality [1–4]. In the field of industrial manufacturing, the industrial metaverse builds a combination of the digital and physical world that highly improves efficiency through engineering and field service. One of the key components of the industrial metaverse is an inspection system that can detect anomalous samples and locate anomalous parts. With the usage of the real-time collected data, a designed algorithm can analyze and give an alarm to the potential risk automatically which greatly saves labor costs and makes the schedule maintenance activities more proactive. Currently, great effort is paid to make use of vision-based techniques like object detection [5] and segmentation [6] to build up such a high-efficiency inspection system, but their limitation is the over-reliance on the large amount of well-labeled anomalous samples which are hard to obtain in practical manufacturing. With the absence of such anomalous data, anomaly detection (AD), also known as one-class classification, has shown great advantages in terms of industrial defect detection and localization, which plays an important role during the development of the industrial metaverse.

Knowledge distillation (KD) [7] using a student-teacher framework has proved to be a feasible paradigm in AD [8–14]. In such a framework, a powerful teacher which is pre-trained on a large-scale dataset, such as ImageNet [15], acts as a feature extractor to extract multi-level features and a naive student is trained to imitate the intermediate features extracted by the teacher. The hypothesis is that since the student network is trained only to mimic the features of normal data extracted from the teacher network, in the test period, it should extract features similar to the teacher network on normal samples and the features extracted by the two networks should be distinguishable enough on those unseen abnormal samples. The similarity between the feature maps at each pixel position can be used as an indicator to predict an anomaly map which can reveal the anomalous area. However, it is always the case that the student-teacher pair tends to extract similar features even on abnormal samples because of the over-generalization of the student network.

We look back to the previous works on S-T models and classify the overall framework into two categories, i.e., the identical S-T network structure and the asymmetric S-T network structure as shown in Fig. 1. In the identical S-T network structure [8, 12] like Fig. 1a, the student network shares completely the same structure as the teacher network. The high similarity of the S-T pair makes it easier to mimic the features extracted by the teacher network on normal samples. However, this may lead to an unwanted over-generalization of abnormal samples because the same network structure is intended to produce the same outcome, which ignores the diversity of abnormal features and makes it hard to tell abnormal areas apart from normal ones. In the asymmetric S-T network structure [13] like Fig. 1b, a smaller student network compared to the teacher network is used for fear that with an entirely identical S-T pair, the features extracted by the two networks may look similar even when given abnormal samples. However, a smaller student network hinders the student's learning ability to mimic the intermediate features of the teacher network during the training period, thus resulting in unsatisfactory prediction results even on normal samples because of poor feature alignment on normal features.

Inspired by previous works, we summarize two key factors that may boost the performance of AD using such an S-T framework. First, S-T networks should extract features of high similarity on normal samples. Second, when encountering abnormal samples, S-T networks should be sensitive to extracting features as different as possible.

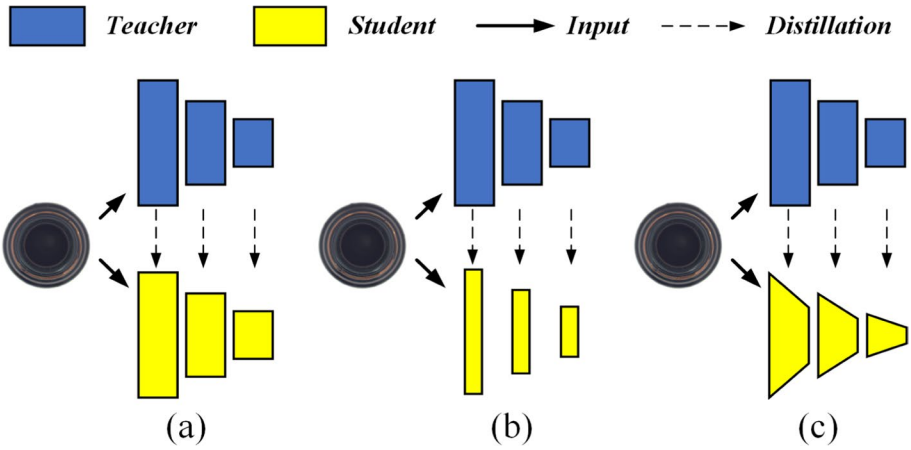

Fig. 1 The overall framework of different kinds of S-T models. **a** Identical S-T network structure. **b**, **c** Asymmetric S-T network structure

That is to say, normal feature alignment and abnormal feature diversity can guarantee a good detection result. Thus, we propose a novel asymmetric S-T framework and introduce a masked feature regeneration (MFR) block to overcome the problems brought by previous KD-based AD methods following the two key factors we summarized above.

Concretely, a novel asymmetric S-T framework is proposed for the diversity of features on the abnormal samples. As discussed above, it is often the case that similar features are extracted even on abnormal samples in the identical S-T network structure. The proposed asymmetric structure of the S-T pair is used so that the model is not intended to produce similar outcomes on abnormal samples. Notably, we build our asymmetric S-T framework in a completely different way from the previous asymmetric S-T network structure, that is, we build the asymmetric S-T pair with an entirely different bottleneck structure like Fig. 1c where distinct features can be extracted from the abnormal samples due to the large structure difference. Specifically, in our asymmetric S-T framework, the WideResNet-like network acts as the teacher, and the ResNext-like network acts as the student. Both student and teacher networks have the same number of network layers while totally different bottleneck block structures. In addition, different from all previous KD-based methods where a naive student is used to learn mimicking features from scratch, we use a student network that is pre-trained on ImageNet [15]. The pre-trained student network is a powerful feature extractor, when facing unseen abnormal samples, it has the potential to extract distinguishable features and ensure the diversity of abnormal features.

Moreover, we draw inspiration from some self-supervised learning methods with masking augmentation [16–19] and propose a masked feature regeneration (MFR) block to ensure a better representation learning process during the training period. The multi-level features extracted by the student network are first masked at random spatial positions and the masked features are then regenerated under the guidance of the teacher network. Such a mask-and-regeneration learning scheme further improves the learning ability of the student network to mimic high-alignment features on normal samples during the training period. In combination with the asymmetric S-T pair and the MFR block, our proposed method AST-MFR achieves competitive performance compared to the state-of-the-art on both anomaly detection and localization tasks on the benchmark dataset MVTec AD [20].

We conclude our main contributions as follows:

- We propose a simple but effective S-T framework for anomaly detection. A novel asymmetric S-T pair is introduced for the diversity of features on the abnormal samples.
- We propose a masked feature regeneration module to further enhance the representation learning ability of the student network to mimic high-alignment features on normal samples.
- Extensive experiments on the MVTec AD dataset demonstrate the effectiveness of our method on both anomaly detection and localization.

2 Related work

AD tasks [21–24], that use only normal data for model training have the ability to distinguish abnormal samples and also locate the anomalous area. In this paper, we focus on anomaly detection and localization of industrial defect images in an unsupervised setting. Different from the supervised method, no other auxiliary information is available except for the images of normal samples. Under such a setting, we broadly categorize the existing methods for AD as reconstruction-based methods, augmentation-based methods, and knowledge distillation-based methods.

2.1 Reconstruction-based methods

In order to learn the distribution pattern of the normal samples, this kind of method completes the task of image reconstruction in an encoding-decoding manner. It is assumed that the model which is only trained to reconstruct normal samples extrapolates badly on abnormal ones and the high reconstruction error can be used to indicate anomalous areas. Some generative models, like autoencoder (AE) [25], variational autoencoder (VAE) [26], and generative adversarial network (GAN) [27] have been widely used as the basic network structure. Yan et al. [28] introduced a multi-level reconstruction framework with an adaptive attention-level transition strategy for anomaly detection and localization. Zhang et al. [29] proposed a multi-task framework to combine image reconstruction with other semantic tasks to learn efficient representations and also introduced a novel hard example mining strategy to further improve the reconstruction quality. Although this kind of method generally works well, the reconstruction error heavily relies on the high quality of reconstructed images thus over-generalization on abnormal areas can be fatal to detecting the defects. In our framework, we distill multi-level knowledge from the teacher network and calculate the anomaly score map in the feature space where abundant information on the anomalies can be obtained compared to the simple reconstruction on the image level. In order to control the over-generalization on abnormal areas, a memory mechanism [30–32] is introduced to AE by building a memory bank to store some prototype features during feature encoding and selectively choose the features in the memory bank to decode. The memory module which linearly combines the stored memory items hinders the reconstruction of some anomaly-like normal areas and requires extra storage space for the established memory bank.

2.2 Augmentation-based methods

Augmentation-based methods try to introduce synthetic, or so-called pseudo outliers in the training period. Schluter et al. [33] created naturally synthetic anomalies by pasting patches of different sizes to different locations of anomaly-free images with Poisson image editing to make the edge of the created anomaly look more natural and more similar to the realistic irregularities. Such a method requires a dedicatedly designed process for creating synthetic anomalies where an extra dataset is needed to serve as the source of the anomaly area images. Moreover, it is also possible that the trained model overfits the synthetic anomalies while performing badly on real-world defects.

2.3 Knowledge distillation-based methods

Knowledge distillation (KD) is first used for model compression [34], where knowledge from an over-parameterized teacher is transferred to a lightweighted student. Recently, KD has been also used for doing AD tasks. A trainable student network is guided to mimic the features on normal samples extracted by a pre-trained teacher network and features on those unseen abnormal samples are expected to be distinct between teacher and student. Deng et al. [11] introduced a reverse distillation paradigm and designed modules to fuse multi-level features and compact embedding space, Cao et al. [12] adopted hard samples mining in training and proposed a novel loss function to distill knowledge from teacher network, Salehi et al. [13] used a smaller student network and distilled multi-level features to detect anomalies. These works all use an identical S-T pair or a smaller student network, and a naive student network trained from scratch, whereas our work introduces a novel asymmetric S-T framework with a pre-trained student, which is more suitable for AD as discussed in Section 1.

3 Proposed framework

This section elaborates on the details of the proposed AST-MFR, whose main framework is illustrated in Fig. 2. The student-teacher framework used in our AST-MFR has a novel asymmetric structure, which is different from the prior works. The teacher network we used is a pre-trained WideResNet-like network, while the student network is a pre-trained ResNext-like network. The two variant versions of ResNet [35], i.e., WideResNet [36] and ResNext [37] have entirely different bottleneck structures. During the training period, multi-level features are extracted by both teacher and student networks respectively. The MFR module is proposed to first mask the features from the student network at each stage and then regenerate the features under the guidance of the teacher network. We calculate the similarity of features at three stages as objective loss and the loss maps are regarded as the anomaly score maps at different resolutions.

3.1 Asymmetric S-T pair

ResNet family models have been widely used in many computer vision tasks for their incredible power in extracting features from the input images. We adopt two different bottleneck structures to build the asymmetric student-teacher network. In Fig. 3, we illustrate three different kinds of bottleneck structures. In the teacher network, we use the bottleneck

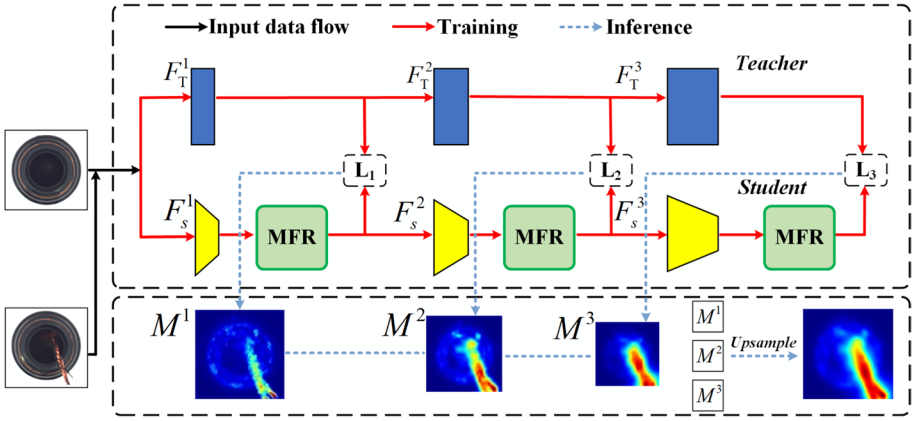

Fig. 2 The overall framework of AST-MFR. The student and teacher with asymmetric structure serve as feature extractors to extract three levels of intermediate features. The features extracted by the student network are first masked at each level and then regenerated under the guidance of the teacher network. At inference time, the similarity between the features extracted by the two networks is calculated and the final anomaly score map is obtained by accumulating the multi-level similarity maps

block like Fig. 3b, which is wider than the traditional bottleneck block used in ResNet, i.e., Fig. 3a. The teacher network itself is expected to extract discriminative enough features that are vital for spotting anomalies. In [36], a network with a wider version of the bottleneck block has been proven to be an effective way to improve performance. Despite that we can further extend the representation capability of the teacher network by simply adding more convolutional layers, a deeper network may face the problem of gradient vanishing if designed improperly. Although ResNet has relieved such a question by introducing residual connection in the bottleneck, it is still possible that only a part of the whole network learns the meaningful representations and the other part contributes little to the final goal.

As for the student network, we want to adopt a bottleneck structure different from the one used in the teacher network, so we turn to the bottleneck block used in ResNext like Fig. 3c. Such a block takes the strategy of repeating layers and uses a factor named cardinality to determine the number of branches with the same structure. The input is first split into several parts and then merged together, and the number of branches is easy to control by setting cardinality to the desired number. This kind of bottleneck has lower complexity than the one used in ResNet while achieving better performance [37].

The whole pipeline of the teacher and student network is demonstrated in Fig. 3d. We follow the overall pipeline of ResNet and take the first three network stages with $4\times$, $8\times$, and $16\times$ downsampling rates while discarding the last stage. The features at higher levels might be too abstract for detecting anomalies because industrial defects usually have little semantic information and some low-level features like edge or outline may be more helpful. Both the student and teacher networks have the same number of layers, while different bottleneck blocks are used in stage 1, stage 2, and stage 3.

3.2 Masked feature regeneration module

Masked image modeling (MIM) is motivated by masked language modeling in the field of NLP [38]. Chen et al. [39] first tried to reconstruct pixels from a masked picture.

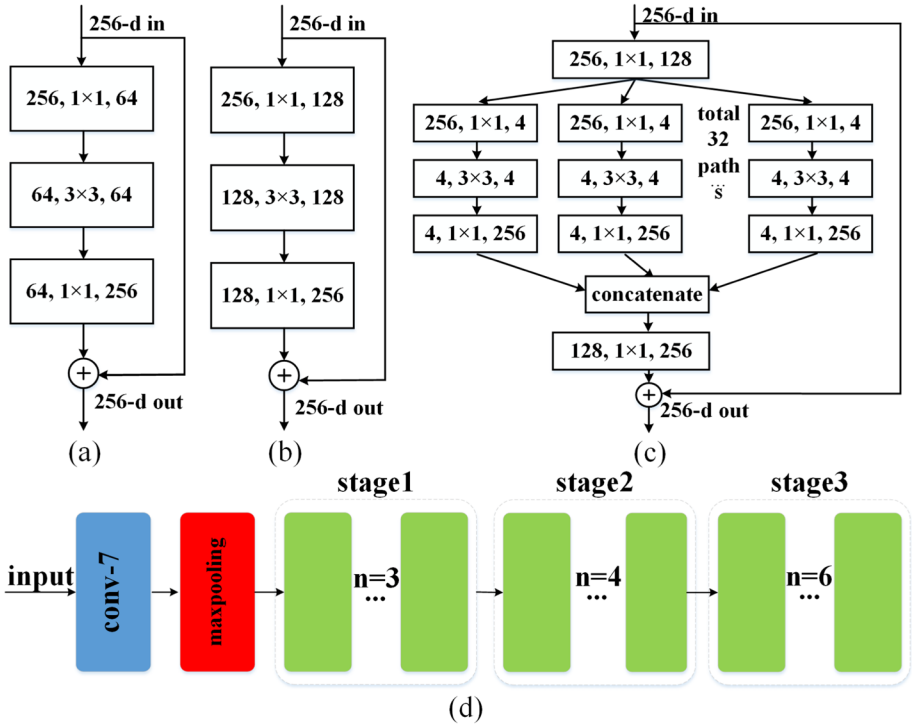

Fig. 3 The bottleneck structure used in the ResNet family models. **a** A bottleneck block in ResNet. **b** A wider bottleneck block in WideResNet. **c** A bottleneck block in ResNext with cardinality = 32. A layer is shown as [input channels, kernel size, output channels]. **d** The pipeline of the student and teacher network. The blue part means the convolutional layer, the red part means the max-pooling layer, the green part means the bottleneck blocks, and n represents the number of blocks used. The teacher and student share the same network pipeline except for the kind of bottleneck used in *stage1* to *stage3*

MIM has proved to be an effective way to learn representations [16], and the key point is that the masking operation can reduce the high spatial redundancy of images. We take inspiration from MIM and further extend it to our work by masking multi-level features of the student network and using the features of the teacher network to guide the regeneration. Specifically, we mask multi-level features instead of the raw input image, and the masked features are regenerated to mimic the corresponding features from the teacher network. In this way, we reformat the knowledge-distilling process as a feature regeneration process under the guidance of the teacher network.

As shown in Fig. 4, the input feature f_i is first masked randomly at the pixel level. Then the masked feature f_m is regenerated by a feature generation block consisting of two convolutional layers. The regenerated feature f_{reg} is then guided to mimic the feature from the teacher network, the detailed process is shown in Algorithm 1. The MFR module is used in all three stages of our student network. With such a feature regeneration operation, the student network can be trained to extract features with higher similarity to the teacher network on normal samples.

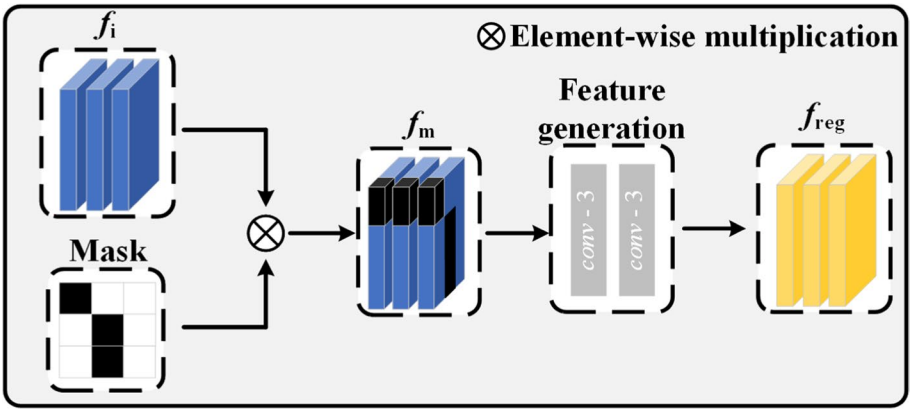

Fig. 4 Illustration of our masked feature regeneration module. A random mask is generated and applied to the input feature. The masked feature is then regenerated by the feature generation block and the regenerated feature is learned to mimic the corresponding feature from the teacher network

3.3 Training process

The training process shifts the feature distribution of the student network on the normal samples to the feature distribution of the teacher network. During the whole training period, the parameters of the teacher network won't be updated and only the student network is trainable. Considering a set of anomaly-free input images $D\{I_1, I_2, \ldots, I_n\}$, for each input image $I_n \in \mathbb{R}^{h \times w \times c}$, where h, w, and c denote the height, the width of the input image, and the number of channels of the input image, respectively, we use the student network and the teacher network to extract the feature at three stages and get the feature outputs $\left\{F_s^1, F_s^2, F_s^3\right\}\left(F_s^n \in \mathbb{R}^{h_n \times w_n \times c_n}\right)$ and $\left\{F_t^1, F_t^2, F_t^3\right\}\left(F_t^n \in \mathbb{R}^{h_n \times w_n \times c_n}\right)$, where n denotes the n-th stage. The loss $\mathcal{L}_{cos}$ and $\mathcal{L}_{L2}$ at position (i, j) can be separately defined as the cosine distance and L2 distance between the normalized feature vectors. We minimize the cosine distance and L2 distance between the feature F_s^n and F_t^n. The cosine distance and L2 distance can be computed respectively by Eq. (1) and Eq. (3). i and j stand for the spatial coordinate of the feature. Particularly, i = 1, 2,..., h_n and j = 1, 2,..., w_n.

$$D_{\cos}^n(i,j) = 1 - \frac{F_s^n(i,j) \odot F_t^n(i,j)}{\|F_s^n(i,j)\|_2 \|F_t^n(i,j)\|_2} \tag{1}$$

$$\mathcal{L}_{\cos} = \sum_{n=1}^{3} \left\{ \frac{1}{h_n \times w_n} \sum_{i,j=0}^{h_n, w_n} D_{\cos}^n(i,j) \right\}, \tag{2}$$

$$D_{L_2}^n(i,j) = \frac{1}{2} \| \frac{F_s^n(i,j)}{\|F_s^n(i,j)\|_2} - \frac{F_t^n(i,j)}{\|F_t^n(i,j)\|_2 \|_{L_2}^2} \tag{3}$$

$$\mathcal{L}_{L_2} = \sum_{n=1}^{3} \left\{ \frac{1}{h_n \times w_n} \sum_{i,j=0}^{h_n, w_n} D_{L_2}^n(i,j) \right\}, \tag{4}$$

Algorithm 1 Masked feature regeneration process

Input: feature input $f_i \in \mathbb{R}^{h \times w \times c}$, masking rate r, feature generation block $G(\cdot)$.

Output: regenerated feature $f_{reg} \in \mathbb{R}^{h \times w \times c}$.

1: Mask = torch.rand(h,w)
2: Mask = torch.where(Mask < r, 0, 1)
3: f_m = torch.mul(f_i, Mask)
4: $f_{reg} = G(f_m)$
5: **return** f_{reg}

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\cos} + \alpha \mathcal{L}_{L_2}. \tag{5}$$

The loss for an entire sample is the average of the feature distance at every pixel position across all three stages as shown in Eq. (2) and Eq. (4). The total loss is the sum of $\mathcal{L}_{\cos}$ and $\mathcal{L}_{L_2}$ shown in Eq. (5). We set $\alpha = 1$ to pay equal attention to the two losses.

3.4 Testing process

During the test period, we terminate the operation of the masking feature but reserve the feature generation blocks. Given a test image I_{test}, we calculate the cosine distance and L2 distance between the feature vectors after the feature generation blocks and the features from the teacher network after the last layer in each stage according to Eq. (1) and Eq. (3). The distances between feature vectors at all pixel positions are calculated to form the distance maps, i.e., M_1, M_2 and M_3 shown in the bottom part of Fig. 2. We obtain the final anomaly score map by accumulating the three maps after upsampling them to the input image size. For anomaly detection, we take the maximum score of the anomaly map to decide whether an image is anomalous.

4 Experimental results

4.1 Experimental protocol

1) Dataset. We assess the performance of our method on the MVTec AD [20] dataset, which is widely used as a benchmark in the field of anomaly detection and localization in an unsupervised setting. MVTec AD comprises 15 different categories, including 10 kinds of objects and 5 kinds of textures. A set of 3629 images is used for training and validation and the other set of 1725 images is used for testing. The test set contains both images with various kinds of defects and anomaly-free images, while the training set contains only images without defects. All the image resolutions vary from 700×700 to 1024×1024 pixels. Pixel-wise ground truth labels are provided for all images with defective areas for the convenience of evaluation.
2) Evaluation metrics. Following the previous works [8, 11, 13, 14, 40–42, 46, 47], we utilize two widely adopted metrics to evaluate the performance of our method on anomaly detection and localization, including the area under the receiver operating characteristic curve (ROCAUC) and the per-region overlap (PRO) curve [9, 43]. ROCAUC is used both for evaluating image-level anomaly detection performance and pixel-level anomaly localization performance. PRO score is derived through the computation of the normalized area under the PRO curve, limited to a 30% pixel-wise false-positive rate, which is a strict metric that weights different-scale anomalies equally.
3) Experimental settings. All images in MVTec AD are re-sized to 256×256. The model is trained to detect and locate anomalies in one category at a time. In our model, we adopt the first three stages of WideResNet50 as the backbone of the teacher network and the first three stages of ResNext50 as the backbone of the student network. Both the student and teacher networks load the pre-trained parameters supplied by the PyTorch. In the training period, we utilize Adam optimizer [44] with $\beta = (0.9, 0.999)$. The learning rate is set to 0.001. We train 200 epochs with a batch size of 32. At test time, a Gaussian filter with $\sigma = 4$ and an average pooling operation with $k = 5$ have been applied to smooth the

final anomaly score map. Our experimental platform is as follows: CPU is Intel Xeon E5-2620, GPU is NVIDIA Titan XP, and PyTorch is configured with Python 3.8.

4.2 Comparison with state-of-the-arts

For anomaly detection, we take ROCAUC as the main evaluation metric and report the performance of our method tested on the MVTec AD dataset as well as the prior state-of-the-art (SOTAs) including, MKD [13], RDAD [11], CutPaste [40], DRAEM [41], PaDiM [42], STFPM [8], DeSTSeg [14], MTHM [29], MLIR [28], ER [46] and OCR [47]. For anomaly localization, we take both ROCAUC and PRO as the evaluation metric and compare the performance of our method with the methods mentioned above (expect for OCR [47] which only reports its image-level anomaly detection performance) and IKD [12] which is specifically designed for anomaly localization.

We validate the performance on the MVTec AD dataset and document the results as shown in Tables 1 and 2. It is worth noting that our method has surpassed the former arts on both image-level detection and pixel-level localization, we achieve a new SOTA performance of 98.7% on the image-level ROCAUC, 98.0% and 94.3% on the pixel-level ROCAUC and PRO. The methods [40] and [41] both use algorithms to introduce synthetic anomalies during the training period, but it is possible that the models overfit the synthetic anomalies and perform weakly on the real anomalies. The method [42] estimates the feature distribution for patch-level Mahalanobis distance, while the method limits the anomaly detection to Mahalanobis distance specific to each patch and it performs subpar in some categories with complex distributions. Noticeably, the methods [8, 11–14] used for comparison also adopt the S-T framework but perform inferior to our method, which is strong evidence to demonstrate the effectiveness of our method. In order to better demonstrate the robustness of our method, we further divide all defective areas into two kinds, one with large scales like broken parts, misplaces, and pollution, and the other with small or slender structures like cracks, missing parts, and stains. As shown in Fig. 5, our method performs well on both kinds of defective areas and predicts the anomaly score maps close to the ground truth mask. To further demonstrate the performance of the model on detecting defects of different scales, we manually divide every defect category into three parts according to the anomalous pixel ratio in all pixels: the defects are defined as small ones if the ratio is under 5%, the defects are defined as middle ones if the ratio is between 5 and 20%, and the defects are defined as large ones if the ratio is higher than 20%. We report the anomaly detection results on defects of different scales in Table 3, and the quantitative results show that our method performs well on all scales of the defects which further demonstrates the robustness of our method.

4.3 Ablation studies

We assess the effectiveness of the AST framework and MFR module on both anomaly detection and localization and report corresponding results in Table 4. We follow the previous work [8] and first use an identical structure of S-T pair as the baseline, where both the student and teacher network adopt the WideResNet50 as the backbone. Then we introduce the MFR module and apply it to the first three stages of the student network. Such a module helps the student to learn features of high alignment on the normal samples and boost the performance on both anomaly detection and localization. We achieve further performance improvement by alternating the whole network by an AST framework, where we adopt

Table 1 Quantitative results of image-level ROCAUC (%) comparison (best scores are highlighted in bold)

	MKD [13]	RDAD [11]	CutPaste [40]	DRAEM [41]	MTHM [29]	PaDiM [42]	MLIR [28]	STFPM [8]	DeSTSeg [14]	ER [46]	OCR [47]	Ours
bottle	99.4	**100.0**	98.3	99.2	99.4	99.9	–	–	–	**100.0**	99.6	**100.0**
cable	89.2	95.0	80.6	91.8	88.4	92.7	–	–	–	98.0	99.1	**99.8**
capsule	80.5	96.3	96.2	98.5	87.1	91.3	–	–	–	98.2	96.2	**98.7**
carpet	79.3	98.9	93.1	97.0	94.3	**99.8**	–	–	–	98.0	99.4	99.2
grid	78.0	**100.0**	99.9	99.9	**100.0**	96.7	–	–	–	99.1	99.6	99.8
hazelnut	98.4	99.9	97.3	**100.0**	96.5	92.0	–	–	–	98.0	98.5	**100.0**
leather	95.1	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	–	–	–	**100.0**	97.1	**100.0**
metal_nut	73.6	**100.0**	99.3	98.7	96.8	98.7	–	–	–	97.1	99.5	**100.0**
pill	82.7	96.6	92.4	98.9	96.4	93.3	–	–	–	**99.0**	98.3	97.1
screw	83.3	97.0	86.3	93.9	92.0	85.8	–	–	–	94.0	**100.0**	95.1
tile	91.6	99.3	93.4	99.6	99.8	98.1	–	–	–	99.6	95.5	**99.9**
toothbrush	92.2	99.5	98.3	**100.0**	99.7	96.1	–	–	–	98.5	98.7	94.4
transistor	85.6	96.7	95.5	93.1	96.9	97.4	–	–	–	**99.5**	98.3	99.0
wood	94.3	99.2	98.6	99.1	98.5	**99.2**	–	–	–	96.6	95.7	98.7
zipper	93.2	98.5	99.4	**100.0**	96.9	90.3	–	–	–	97.8	99.0	98.4
average	87.8	98.5	95.2	98.0	96.2	95.4	90.4	95.2	98.6	98.2	98.3	**98.7**

MLIR, STFPM and DeSTSeg only report their average results on the dataset in their original papers

Table 2 Pixel-level ROCAUC (%) and PRO (%) results comparison (best scores are highlighted in bold)

	MKD [13]	RDAD [11]	CutPaste [40]	DRAEM [41]	PaDiM [42]	STFPM [8]	DeSTSeg [14]	IKD [12]	MTHM [29]	MLIR [28]	ER [46]	Ours
average	90.9/-	97.8/93.9	96.0/-	97.3/-	97.5/92.1	97.1/92.4	97.9/-	97.8/92.3	97.4/-	96.9/94.0	97.9/92.7	**98.0/94.3**

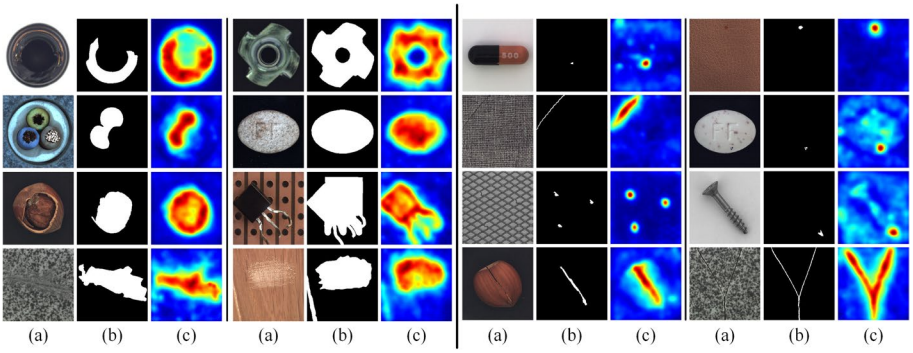

Fig. 5 Visualization examples of large defects (left) and tiny and indistinctive defects (right). For each example: **a** source image, **b** ground truth mask, **c** prediction of anomaly score map

Table 3 Quantitative results of image-level ROCAUC (%) on defects of different scales

Category	Small	Middle	Large
bottle	100.0	100.0	100.0
cable	98.9	99.0	100.0
capsule	97.8	99.7	100.0
carpet	98.1	100.0	100.0
grid	99.7	100.0	–
hazelnut	100.0	100.0	100.0
leather	100.0	100.0	–
metal_nut	100.0	100.0	100.0
pill	96.0	99.1	100.0
screw	94.1	–	–
tile	100.0	100.0	99.9
toothbrush	89.6	100.0	100.0
transistor	100.0	100.0	96.3
wood	96.7	99.7	98.0
zipper	97.6	99.3	100.0
average	97.9	99.8	99.5

Table 4 Effects of AST and MFR on the final ROCAUC results

Exp.	MFR	AST	Image-level	Pixel-level
1			96.0	97.0
2		√	96.5	97.3
3	√		98.4	97.9
4	√	√	**98.7**	**98.0**

the pre-trained WideResNet50 as the backbone of the teacher network and the pre-trained ResNext50 as the backbone of the student network. Such an asymmetric structure and a more experienced student benefit a lot for the insurance of the diversity of the features on abnormal features as well as a better knowledge transfer process during the training period.

Table 5 Complexity of different backbones used for the student network

Backbone	Para(M)	FLOPs(G)	Time(ms)	Image-level	Pixel-level
WideResNet50	49.6	26.6	9.6	98.5	97.9
ResNext50	33.2	18.9	10.1	98.7	98.0
ResNet50	36.3	21.2	8.9	98.5	98.0

Table 6 Effects of the asymmetric S-T structure on the final ROCAUC results

Teacher	ResNet50	WideResNet50	ResNext50	WideResNet50
Student	ResNet50	WideResNet50	ResNext50	ResNext50
Image-level	97.9	98.4	98.3	**98.7**
Pixel-level	97.7	97.9	97.8	**98.0**

Complexity analysis We study the computing complexity of the proposed method. The WideResNet50 is chosen as the teacher network and different backbones are chosen as student networks. As shown in Table 5, we report three metrics used for measuring the complexity of the model: Parameters, FLOPs, inference time, and the corresponding performance of different student networks. Although it costs more inference time when using ResNext50 as the backbone, it has both fewer parameters and FLOPs compared to WideResNet50 and ResNet50. What's more, the best performance is achieved on both tasks of anomaly detection and localization with ResNext50. Therefore, we choose to use ResNext50 as the final backbone of the student network for the better balance between model complexity and model performance.

Effects of asymmetric S-T structure We further conduct experiments to explore the necessity of using an asymmetric S-T structure. We retrain the model with three S-T networks of the same structure (with an MFR module) and compare the model performance with the one using an asymmetric S-T structure. As the results are shown in Table 6, the models with the same S-T structure like ResNet50, WideResNet50, and ResNext50 all perform inferior to the one with an asymmetric S-T structure. Since WideResNet50 and ResNext50 are strong feature extractors as teacher networks and have good learning ability as student networks, they have better performance than ResNet50 on both anomaly detection and localization. While we further boost performance by introducing the asymmetric S-T structure, as discussed before, the novel structure further enlarges the feature difference in the anomalous areas. The experimental results validate the effectiveness and rationality of the asymmetric S-T structure used in our method.

Effects of different masking rates We investigate the effects of masking rates on the final results. As shown in Fig. 6, we choose different masking rates r = {0.1, 0.3, 0.5, 0.7, 0.9} and report the following results on both anomaly detection and localization. We achieve the optimal result on the masking rate of 0.3 and 0.5. The key point behind such an MIM method is the reduction of spatial redundancy of the feature map and a low masking rate like 0.1 does not achieve the desired effect. On the other hand, the excessive operation of

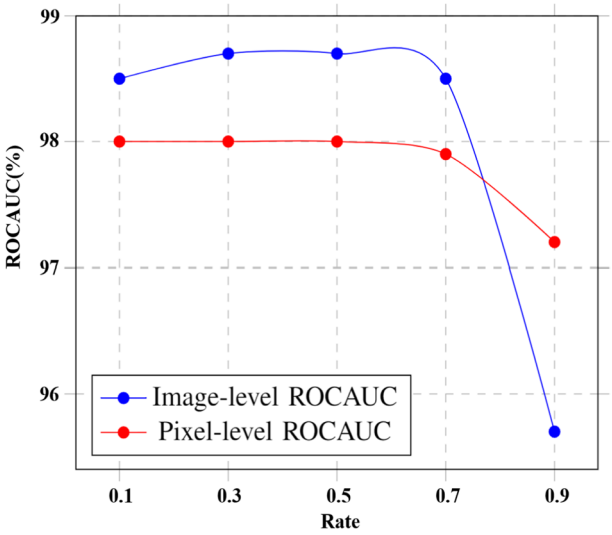

Fig. 6 Model performance under different masking rates. We depict the trend of variation of image-level ROCAUC in blue dots, and the one of pixel-level ROCAUC in red dots

masking also harms representation learning since the generation blocks may find it hard to regenerate features with little useful information.

Effectiveness of a pre-trained student network We evaluate the performance of our proposed method with a student network trained from scratch and display the results in the second column of Table 7. A more experienced student network has a stronger capability of representation learning during the training period. It extracts more discriminative features on ab-normal samples compared to a naive student network trained from scratch, which contributes a lot to both anomaly detection and localization.

Combinations of different backbone usage We use three kinds of different backbones to get six different combinations of asymmetric S-T pairs and assess the corresponding performance on both anomaly detection and localization as shown in the last six columns of Table 7. An obvious performance drop happens when ResNet50 acts as the teacher network for its lack of representative capacity. We find that even with a simpler student like ResNet50, the pre-trained student combined with the MFR module still has a strong ability to perform feature learning and gets satisfactory performance. The optimal result is achieved under the combination of WideResNet50 and ResNext50 due to the large architecture gap in terms of the bottleneck block.

Effects of different locations to add MFR module In this work, we propose the MFR module to mask the student's features first and then regenerate the features according to the teacher's features. Since there are three different stages during the feature extraction, we conduct experiments to study the effect of the location of the added MFR module on the final results. As the results shown in Table 8, we observe a performance boost when singly adding the MFR module to each stage, and the MFR module added at stage2 helps most since the middle-level features are not that abstract while containing information like edges

Table 7 Effects of a pre-trained(P) student and different backbones on the final ROCAUC results

Teacher	WideResNet50	ResNet50	WideResNet50	WideResNet50	ResNext50	ResNext50	ResNet50
Student	ResNext50	WideResNet50(P)	ResNext50(P)	ResNet50(P)	WideResNet50(P)	ResNet50(P)	ResNext50(P)
Image-level	98.4	98.1	**98.7**	98.5	98.6	98.4	98.1
Pixel-level	97.9	97.9	**98.0**	98.0	98.0	98.0	97.9

Table 8 Effects of different location to add MFR module on the final ROCAUC results

stage1	stage2	stage3	Image-level	Pixel-level
			96.0	97.0
√			97.4	97.3
	√		98.3	97.8
		√	98.1	97.6
√	√		98.6	97.8
√		√	98.4	97.9
	√	√	98.5	98.0
√	√	√	**98.7**	**98.0**

and contours which are beneficial for detecting anomalies. We observe further performance boost when adding the MFR module to the different stages at the same time and the best performance is achieved by adding the MFR module to all three stages. The results have shown that the designed MFR module can strengthen the student network's learning ability and the features of all three stages are vital for anomaly detection.

4.4 Failure cases and analyses

Although our method achieves good performance on most of the categories in the MVTec AD dataset, there is a significant accuracy drop in screw and toothbrush. We visualize some of the false detection examples and try to make an explanation for such failures. As shown in Fig. 7, we find that our method is sensitive to some disturbances like fiber and stain existing in the background areas and predicts a relatively high anomaly score in these areas. Since we take the maximum of the anomaly score map as the anomaly score of the given image, such a normal image without other kinds of real defects is misclassified as an anomaly due to the occurrence of a high anomaly score in the background area. It is reasonable, however, since the model is trained in the absence of abnormal samples and does not have any prior knowledge of the real defects, it is only taught to mimic features from the teacher network of the normal areas. Such disturbance, although regarded as a pseudo flaw, does never appear in the training set, so it is excusable that our method treats it as irregularity. In order to make up for such a mistake, we make use of SAM [45] to get the binary mask of the foreground object and multiply the mask with the anomaly score map to restrict the predicted anomalous areas only appearing around the object area. After removing the disturbance from the background (RB), we further boost image-level ROCAUC on both screw and toothbrush by 2.1% and 5.9% as shown in Table 9. Although there is a performance drop in transistor where background areas are not easy to separate in some cases, we improve the performance on most of the object categories after removing disturbance in the background. A more effective way to relieve such unwanted disturbance in the background areas can be further explored in future research for even better anomaly detection performance.

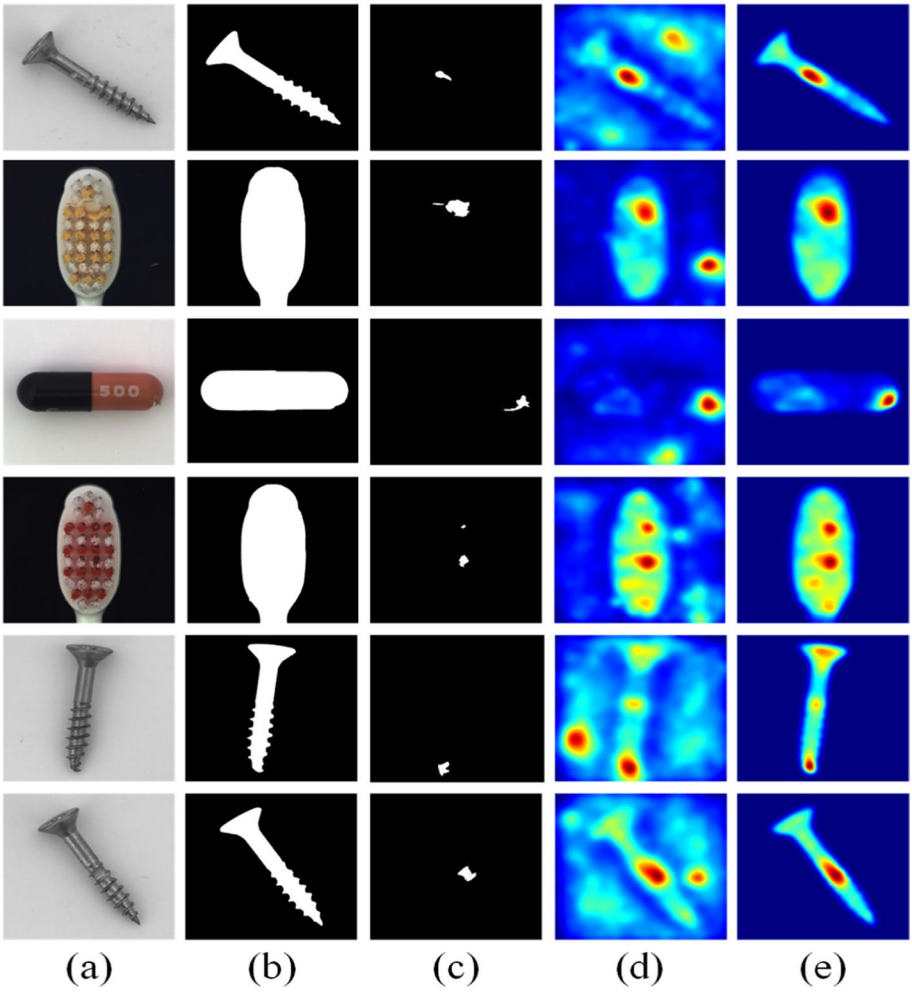

Fig. 7 Failure examples of our method. For each example: **a** source image, **b** binary mask of the foreground object, **c** ground truth of the anomaly area, **d** prediction before removing background areas, **e** prediction after removing background areas

5 Conclusion

In this paper, we follow the KD-based methods and propose AST-MFR for AD. We discuss the shortcomings of the previous KD-based methods for AD and summarize two key factors towards a better performance of AD: normal feature alignment and abnormal feature diversity. According to this, we propose a novel S-T framework to complete the task of AD in an unsupervised setting, and our asymmetric structure of student and teacher networks brings a better model performance. In addition, we further extend the method of MIM and propose the MFR module to mask the features of the student network and regenerate them under the guidance of the teacher network. Such

Table 9 Effects of removing the background (RB) on the final image-level ROCAUC results

	w/o RB	w/ RB
bottle	100.0	100.0
cable	99.8	99.9
capsule	98.7	99.2
hazelnut	100.0	100.0
metal_nut	100.0	100.0
pill	97.1	97.1
screw	95.1	97.1
toothbrush	94.4	100.0
transistor	99.0	98.6
zipper	98.4	98.4
average	98.3	99.0

a feature regeneration process enhances the learning ability of the student network for better knowledge transfer and results in a high similarity of features on the normal samples. Experimental results show that our method has outperformed the prior arts on both anomaly detection and localization.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 62171269, and in part by the China Postdoctoral Science Foundation under Grant 2022M722037.

Data availability Data will be made available on reasonable request.

Declarations

Conflicts of interests The authors declare that they have no conflict of interest.

References

1. Wali A, Lisan A, Ather H, Qasim M, Abid MU (2023) Application in multimedia: from camera to VR. Multimedia Tools Appl 82(8):11721–11751. https://doi.org/10.1007/s11042-022-13687-1
2. Al-Amri S, Hamid S, Noor NFM, Gani A (2023) A framework for designing interactive mobile training course content using augmented reality. Multimedia Tools Appl 82(20):30491–30541. https://doi.org/10.1007/s11042-023-14561-4
3. Wu K, Yang Y, Liu Q, Zhang X-P (2023) Focal stack image compression based on basis-quadtree representation. IEEE Trans Multimedia 25:3975–3988. https://doi.org/10.1109/TMM.2022.3169055
4. Wu K, Yang Y, Liu Q, Jiang G, Zhang X-P (2023) Hierarchical independent coding scheme for varifocal multiview images based on angular-focal joint prediction. IEEE Trans Multimedia: 1–13. https://doi.org/10.1109/TMM.2023.3306072
5. Ashiba HI, Ashiba MI (2023) Novel proposed technique for automatic fabric defect detection. Multimedia Tools Appl 82(20):30783–30806. https://doi.org/10.1007/s11042-023-14368-3
6. Cheng L, Yi J, Chen A, Zhang Y (2023) Fabric defect detection based on separate convolutional unet. Multimedia Tools Appl 82(2):3101–3122. https://doi.org/10.1007/s11042-022-13568-7
7. Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531. https://doi.org/10.48550/arXiv.1503.02531
8. Wang G, Han S, Ding E, Huang D (2021) Student-teacher feature pyramid matching for anomaly detection. In: Proceeding of the British Machine Vision Conference, p 306
9. Bergmann P, Fauser M, Sattlegger D, Steger C (2020) Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: Proceeding of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition, pp 4183–4192. https://doi.org/10.1109/CVPR42600.2020.00424
10. Yamada S, Hotta K (2021) Reconstruction student with attention for student teacher pyramid matching. arXiv preprint arXiv:2111.15376. https://doi.org/10.48550/arXiv.2111.15376
11. Deng H, Li X (2022) Anomaly detection via reverse distillation from one-class embedding. In: Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9737–9746. https://doi.org/10.1109/CVPR52688.2022.00951
12. Cao Y, Wan Q, Shen W, Gao L (2022) Informative knowledge distillation for image anomaly segmentation. Knowl-Based Syst 248:108846. https://doi.org/10.1016/j.knosys.2022.108846
13. Salehi M, Sadjadi N, Baselizadeh S, Rohban MH, Rabiee HR (2021) Multiresolution knowledge distillation for anomaly detection. In: Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14902–14912. https://doi.org/10.1109/CVPR46437.2021.01466
14. Zhang X, Li S, Li X, Huang P, Shan J, Chen T (2023) DeSTSeg: Segmentation guided denoising student-teacher for anomaly detection, in: Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3914–3923. https://doi.org/10.1109/CVPR52729.2023.00381
15. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: A large-scale hierarchical image database. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848
16. He K, Chen X, Xie S, Li Y, Doll´ar P, Girshick R (2022) Masked autoencoders are scalable vision learners. In: Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 16000–16009. https://doi.org/10.1109/CVPR52688.2022.01553
17. Wei C, Fan H, Xie S, Wu C-Y, Yuille A, Feichtenhofer C (2022) Masked feature prediction for self-supervised visual pre-training. In: Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14668–14678. https://doi.org/10.1109/CVPR52688.2022.01426
18. Radosavovic I, Xiao T, James S, Abbeel P, Malik J, Darrell T (2023) Real world robot learning with masked visual pre-training. In: Proceeding of the Conference on Robot Learning, pp 416–426
19. Tao C, Zhu X, Su W, Huang G, Li B, Zhou J, Qiao Y, Wang X, Dai J (2023) Siamese image modeling for self-supervised vision representation learning. In: Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2132–2141. https://doi.org/10.1109/CVPR52729.2023.00212
20. Bergmann P, Fauser M, Sattlegger D, Steger C (2019) MVTec AD-A comprehensive real-world dataset for unsupervised anomaly detection. In: Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9592–9600. https://doi.org/10.1109/CVPR.2019.00982
21. Liu L, Zhang J, Fu X, Liu L, Huang Q (2019) Unsupervised segmentation and elm for fabric defect image classification. Multimedia Tools Appl 78(9):12421–12449. https://doi.org/10.1007/s11042-018-6786-7
22. Wan D, Gao C, Zhou J, Shen X, Shen L (2023) Unsupervised fabric defect detection with high-frequency feature mapping. Multimedia Tools Appl. https://doi.org/10.1007/s11042-023-16340-7
23. Jiang W, Yang K, Qiu C, Xie L (2023) Memory enhancement method based on skip-ganomaly for anomaly detection. Multimedia Tools Appl. https://doi.org/10.1007/s11042-023-16317-6
24. Sharma P, Gangadharappa M (2023) An attention-augmented driven modified two-fold u-net anomaly detection model for video surveillance systems. Multimedia Tools Appl. https://doi.org/10.1007/s11042-023-16728-5
25. Rudolph M, Wandt B, Rosenhahn B (2019) Structuring autoencoders. In: Proceeding of the IEEE/CVF International Conference on Computer Vision Workshops, pp 615–623. https://doi.org/10.1109/ICCVW.2019.00075
26. Kingma DP, Welling M (2014) Auto-encoding variational bayes. In: International Conference on Learning Representations
27. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2020) Generative adversarial networks. Commun ACM 63(11):139–144. https://doi.org/10.1145/3422622
28. Yan Y, Wang D, Zhou G, Chen Q (2021) Unsupervised anomaly segmentation via multilevel image reconstruction and adaptive attention-level transition. IEEE Trans Instrum Meas 70:1–12. https://doi.org/10.1109/TIM.2021.3107586.doi:10.1109/TIM.2021.3107586
29. Zhang C, Wang Y, Tan W (2023) MTHM: Self-supervised multitask anomaly detection with hard example mining. IEEE Trans Instrum Meas 72:1–13. https://doi.org/10.1109/TIM.2023.3276529.doi:10.1109/TIM.2023.3276529
30. Gong D, Liu L, Le V, Saha B, Mansour MR, Venkatesh S, Hengel Avd (2019) Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceeding of the IEEE/CVF International Conference on Computer Vision, pp 1705–1714. https://doi.org/10.1109/ICCV.2019.00179

31. Hou J, Zhang Y, Zhong Q, Xie D, Pu S, Zhou H (2021) Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In: Proceeding of the IEEE/CVF International Conference on Computer Vision, pp 8791–8800. https://doi.org/10.1109/ICCV48922.2021.00867
32. Tian Y, Pang G, Liu Y, Wang C, Chen Y, Liu F, Singh R, Verjans JW, Wang M, Carneiro G (2023) Unsupervised anomaly detection in medical images with a memory-augmented multi-level cross-attentional masked autoencoder. In: Machine Learning in Medical Imaging, pp 11–21. https://doi.org/10.1007/978-3-031-45676-3_2
33. Schluter HM, Tan J, Hou B, Kainz B (2022) Natural synthetic anomalies for self-supervised anomaly detection and localization, in: Proceeding of the European Conference on Computer Vision, pp 474–489. https://doi.org/10.1007/978-3-031-19821-2_27
34. Buciluˇa C, Caruana R, Niculescu-Mizil A (2006) Model compression. In: Proceeding of the ACM Conference on Knowledge Discovery and Data Mining, pp 535–541. https://doi.org/10.1145/1150402.1150464
35. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778. https://doi.org/10.1109/CVPR.2016.90
36. Zagoruyko S, Komodakis N (2016) Wide residual networks. In: Proceeding of the British Machine Vision Conference
37. Xie S, Girshick R, Doll´ar P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1492–1500. https://doi.org/10.1109/CVPR.2017.634
38. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I et al (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9
39. Chen M, Radford A, Child R, Wu J, Jun H, Luan D, Sutskever I (2020) Generative pretraining from pixels. In: Proceeding of the International Conference on Machine Learning, pp 1691–1703
40. Li C-L, Sohn K, Yoon J, Pfister T (2021) CutPaste: Self-supervised learning for anomaly detection and localization. In: Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9664–9674. https://doi.org/10.1109/CVPR46437.2021.00954
41. Zavrtanik V, Kristan M, Skoˇcaj D (2021) DRÆM–A discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceeding of the IEEE/CVF International Conference on Computer Vision, pp 8330–8339. https://doi.org/10.1109/ICCV48922.2021.00822
42. Defard T, Setkov A, Loesch A, Audigier R (2021) Padim: a patch distribution modeling framework for anomaly detection and localization. In: Proceeding of the IEEE International Conference on Pattern Recognition, pp 475–489. https://doi.org/10.1007/978-3-030-68799-1_35
43. Bergmann P, Löwe S, Fauser M, Sattlegger D, Steger C (2019) Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In: Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp 372–380. https://doi.org/10.5220/0007364503720380
44. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. In: International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1412.6980
45. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W-Y, Doll´ar P, Girshick RB (2023) Segment anything, arXiv preprint arXiv:2304.02643. https://doi.org/10.48550/arXiv.2304.02643
46. Zhu Z, Chen S, Huang Y, Leng T (2023) Enhancing industrial anomaly detection using edge image reconstruction with neighbor masked convolutional transformer block. In: International Conference on Intelligent Computing and Human-Computer Interaction, pp 372–376. https://doi.org/10.1109/ICHCI58871.2023.10277867
47. Liang Y, Zhang J, Zhao S, Wu R, Liu Y, Pan S (2023) Omni-frequency channel-selection representations for unsupervised anomaly detection. IEEE Trans Image Process 32:4327–4340. https://doi.org/10.1109/TIP.2023.3293772