Complex & Intelligent Systems (2024) 10:4853–4865
https://doi.org/10.1007/s40747-024-01412-4

ORIGINAL ARTICLE

Dual-student knowledge distillation for visual anomaly detection

Jutao Hao[1] · Kai Huang[2] · Chen Chen[3] · Jian Mao[2]

Received: 16 August 2023 / Accepted: 9 March 2024 / Published online: 9 April 2024

Abstract

Anomaly detection poses a significant challenge in the industry and knowledge distillation constructed using a frozen teacher network and a trainable student network is the prevailing approach for detecting suspicious regions. Forward and reverse distillation are the main ways to achieve anomaly detection. To design an effective model and aggregate detection results, we propose a dual-student knowledge distillation (DSKD) based on forward and reverse distillation. Taking advantage of the priority of reverse distillation to obtain high-level representation, we combine a skip connection and an attention module to build a reverse distillation student network that simultaneously focuses on high-level representation and low-level features. DSKD uses a forward distillation network as an auxiliary to allow the student network to preferentially obtain the query image. For different anomaly score maps obtained by the dual-student network, we use synthetic noise enhancement in combination with image segmentation loss to adaptively learn the weight scores of individual maps. Empirical experiments conducted on the MVTec dataset show that the proposed DSKD method achieves good performance on texture images as well as competitive results on object images compared with other state-of-the-art methods. Meanwhile, ablation experiments and a visualization analysis validate the contributions of each of the model's components.

Keywords Anomaly detection · Knowledge distillation · Dual-student

Introduction

Anomaly detection is a critical process that involves determining whether a given sample deviates from the normal distribution and detecting its unusual components, and it has a wide range of applications in industrial control [12, 33], product quality control [4, 43], and other fields [5, 22, 32]. Real-world datasets present challenges due to their characteristics of widely varying distributions and the scarcity of anomalous samples, which result in limited prior knowledge about an anomalous class. For single-class detection, such as in singular value decomposition [17] and one-class SVM [16], a feature is mapped to higher dimensions to increase the characterizability of normal samples; however, this is accompanied by relatively limited feature extraction capabilities. In image-related fields, deep neural networks have shown promising results. Methods like [37] extract latent features that characterize the dataset while detecting objects. However, over-parameterization often occurs for networks trained only on normal samples, and how to apply the feature extraction capability of deep neural networks is a hot topic [30, 31].

The teacher–student (T–S) architecture [41] is an important part of anomaly detection, and for situations where only normal samples can be observed, anomalies will be identified by observing a teacher model trained with a large amount of data and comparing it with a student model trained with normal samples. In contrast to the traditional T–S architecture that pursues smaller student network parameters to achieve fast inference, the T–S architecture for anomaly detection exploits the inconsistency of different networks on the training data to detect anomalous samples. The teacher network's

✉ Kai Huang
kaihuang@jmu.edu.cn

Jutao Hao
haojt@sdju.edu.cn

Chen Chen
chchen@tongji.edu.cn

Jian Mao
maojian@jmu.edu.cn

[1] School of Electric Information Engineering, Shanghai Dianji University, Shuihua Rd, Shanghai 201306, China

[2] College of Computer Engineering, Jimei University, Yinjiang Rd, Xiamen 361021, Fujian, China

[3] College of Electronic and Information Engineering, Tongji University, Cao'an Rd, Shanghai 201804, China

parameters are typically trained on large datasets, such as ImageNet [15], to obtain semantically strong descriptors. Meanwhile, the student network takes normal samples as input during training, compares its feature map with the corresponding hierarchical teacher network's map, and uses the cosine similarity or mean squared error as loss. The assumption of the T–S architecture is that the ability of the teacher network to encode the images is transferred to the student network as much as possible. The student network is trained with normal samples by default, which differs from the pre-trained teacher network especially when dealing with anomaly images. Anomaly score maps result from comparing the feature maps of the two networks.

Since the student network uses the teacher network as the learning object, constructing the student network more efficiently is a key concern for anomaly detection. Some studies have approached the construction of different pre-trained teacher models or distillation methods. For instance, Xu et al. [46] proposes to use multiple teacher networks to handle different detection objects while Deng and Li [14] proposes to use reverse distillation to obtain comprehensive representations of teacher networks. Jiang et al. [23] conduct work based on [14] and introduce pixel-level and feature-level masking to alleviate the overgeneralization problem. Cao et al. [8] argues that increasing the knowledge of the student network is an effective way to improve model recognition. In previous studies, which have focused primarily on guided learning of single student networks, including the use of methods such as soft logit [46] and one-class embedding [14]. Ma et al. [29] in showing the effectiveness of training student networks using multiple pre-trained teacher models. Since different distillation methods have various advantages, it is beneficial to construct diverse student networks and aggregate their score maps.

To construct student networks that effectively generate recognition score maps, we adopt a dual-student approach that leverages high-level and low-level representations and performs subsequent aggregation. We propose a student network architecture for anomaly detection based on a skip connection and an attention mechanism based on reverse distillation. The attention mechanism supporting the student network determines which feature maps are more important in the teacher's hierarchy. In addition, we develop a forward distillation student network that integrates the anomaly score maps obtained from both students using synthetic noise. Our main contributions are given below.

- We propose a novel multi-student knowledge distillation framework for anomaly detection and localization named DSKD. Through synthetic noise, DSKD aggregates the score maps obtained from two different student networks, thus leading to a more powerful representation of learning.
- To further improve the efficiency and effectiveness of anomaly detection and localization during reverse distillation, we propose a skip connection architecture to help the student network obtain the information of the layer corresponding to the teacher; moreover, an attention module is added to help the student recombine the features.

The remainder of this paper is divided into five sections. "Related work" presents related work on anomaly detection and knowledge distillation. "Mythology" introduces our proposed method. "Experimental results and analysis" presents our experiments and analysis. A conclusion is given in "Conclusion".

Related work

Anomaly detection

Anomaly detection, also known as outlier detection or novelty detection, involves identifying samples that deviate from the rest of the observations. In our work, we assume that a model is trained using only normal samples, and since there is no supervisory information from other classes, the problem is treated as novelty detection. Traditional methods such as singular value decomposition and one-class SVM construct a hypersphere or hyperplane to check whether outliers are far from the hypersphere center. Deep learning-based methods are also used in anomaly detection due to the effectiveness of neural networks in feature extraction.

Methods based on autoencoders utilize reconstruction error as a primary measure for judging sample abnormality. This approach has found wide application in various domains such as video analysis [28], Internet of Things (IoT) [48], and railway turnout inspection [9]. For instance, [6] evaluates the application of an autoencoder for visual fault detection and finds that there are deficiencies in the reconstruction of high-frequency textures and small details using a convolutional autoencoder combining $l2$ loss and the structural similarity index. Üzen et al. [40] combines a convolution layer and swin transformer, where the former provides spatial properties and the latter provides global semantic properties. Meanwhile, [25] reduces false positives of an autoencoder by contrastive learning of complex shapes, sizes, and colors of the recruitment samples. In addition, skip connections are widely used in autoencoders to help the model reconstruct sharpness while preserving both high- and low-frequency information [11].

GAN-based methods use normal samples to train a model, and then they compare the generator, discriminator, and reconstructor for detection [45]. AnoGAN [36] uses normal data to train the model and compute the reconstruction errors between the generator and discriminator. Moreover, GANomaly [2] uses latent vectors of generators and discriminators to obtain an anomaly score. Skip-GANomaly [3] combines reconstruction metrics and latent representations using skip connections to improve the underdetection of small-scale anomalies in GANomaly. The discriminator identifies deviations from the normal distribution by comparing the reconstructed image with the original, particularly when the reconstruction is of poor quality. However, relying solely on reconstruction errors means that we do not fully utilize the potential of a large dataset. To address this limitation, a teacher–student (T–S) architecture was introduced. Furthermore, it was found that skip connections majorly improve anomaly detection as they provide high-frequency features as well as relevant support for our network's design.

Knowledge distillation

With the advent of deep learning, networks designed to enhance classification accuracy have become increasingly deep and wide, resulting in substantial computational requirements for both training and testing. Conventional knowledge distillation is designed to give models a lighter architecture and reduce inference time. Knowledge distillation techniques involve leveraging a trained teacher network to transfer knowledge to a student network through soft loss distillation, thereby offering the potential for real-time predictions. Bergmann et al. [7] proposes a T–S network trained on large datasets and normal samples; anomaly detection involves processing external datasets by comparing the two networks. To further improve the student network's ability to recover the original image, a denoising process [49] is introduced. It trains the student network with an anomaly mask to recover the original image using synthetic corrupted normal images. Under the assumption that the teacher model is underutilized, a gradient-based adaptive anomaly localization approach [35] based on the distillation of intermediate layer feature maps is proposed. Maps are sufficiently distilled to simplify the model and bypass the patching process. Tong et al. [39] is similar to [50] in that they both introduce self-supervised mask training for reinforcement learning of WideResNet50-based single-class prototype models as well as employ a feature diffusion module to identify large-area anomalies. Ma et al. [29] argues that teacher networks with different structures or initial parameters provide features from different perspectives and thus also supports the notion that constructing reliable feature maps is important in T–S-based anomaly detection.

Mythology

In anomaly detection, training data typically consists of only normal images. A crucial aspect of generating anomaly score maps involves constructing comparison networks based on teacher networks. There is a need to improve the sensitivity of students to the semantics and details to achieve both sample- and pixel-level anomaly detection. In our proposed method, we employ a teacher network as an anchor point and utilize distillation in multiple directions to obtain two student networks. Subsequently, we design a fusion network to combine the anomaly score maps generated by these students. The validity of our approach is demonstrated through both semantic and pixel anomaly detection.

Problem definition

An important assumption for anomaly detection is that anomalous samples are difficult to acquire or observe during training. Consistent with the literature [14] on anomaly detection, we let $L_t = \{L_t^1, ..., L_t^n\}$ be a normal sample in the dataset that only appears in training, thereby allowing the model to learn the normal distribution. Let $L_q = \{L_q^1, ..., L_q^n\}$ denote the samples to be detected; this set contains normal and abnormal data. Our goal is to construct a model trained with L_t that correctly identify the samples in L_q. Normal samples in L_t and L_q conform to the same distribution, and samples outside of this distribution are considered anomalies. Taking the screw category of samples as an example, the model trained with screw samples is only used for detection within that category. Data that deviates from the distribution of that category will be considered an anomaly.

Network architecture

Figure 1 shows our model architecture. Unlike those with multiple teachers [29], our architecture uses only a pretrained teacher network E that extracts multi-scale representations of normal samples. Since anomaly detection is usually pixel-level mining, i.e., it is similar to image segmentation, it is sensitive to both high-level and low-level representation details. We propose to extract these representations using students D_f and D_r. D_r uses reverse distillation to obtain high-level representations while connecting E's previous feature maps to get different levels representations. Meanwhile, we retain low-level representations using forward distillation network D_f and fuse the D_f network with the D_r network. In the inference stage, multi-scale anomaly detection is performed using D_r and D_f, separately, and a unique anomaly map is obtained by each component of the fused network.

Since we focus on the network design of the student network, for the teacher network we simply use WideResNet50: it has more capacity and benefits less from input repetition. The parameters of the teacher network are learned from the ImageNet and are fixed. According to the network design of WideResNet50, the teacher network is divided into four network structures, with E_1–E_4 denoting the blocks from largest to smallest. Correspondingly, the network structures of D_{f} and D_{r}, which we will explain in "Forward distillation student network" and "Reverse distillation student network", respectively, are similar to those of the teacher's blocks. The adaptive fusion of D_{f} and D_{r} is elaborated on in "Anomaly score map fusion".

Forward distillation student network

The teacher network learns about a diverse and large amount of external data by pre-training, and the student network focuses on the normal samples of the dataset during its learning process. We examine the inconsistencies between the feature maps of the teacher and student networks to mine anomaly samples with query data. Forward distillation inputs images to the teacher and student network to obtain feature maps z^t and z^f, respectively. By cosine similarity loss, keeping z^t and z^f in the same direction, the student feature map is kept numerically diverse. The loss is shown in Eq. (1).

$$\mathrm{loss} = 1 - \frac{z^t \cdot z^f}{\|z^t\|_2 \cdot \|z^f\|_2} \tag{1}$$

$\|z^t\|_2$ and $\|z^f\|_2$ are $l2$-norms of z^t and z^f. In the design of the D_{f} network, we use a simple approach of directly feeding data to the teacher and student. We expect the student network to directly use the data to obtain a more detailed feature map. At the same time, by not making additional connections between the student and teacher, we prevent the teacher from copying parameters to the student network and hence keep the student network sensitive to anomalies.

In a prediction phase, the anomaly map calculates the cosine similarity of each z^f to the corresponding z^t through Eq. (1). Position-by-position difference results are recorded and expanded to the original size by interpolation. In this process, we return the first three anomaly maps to be fused with the reverse distillation results.

Reverse distillation student network

While D_{f} gives a superficial representation of an image, in the case of anomalous samples being outside of the normal distribution, the results of D_{f} and E converges due to their use of the same data. Deng and Li [14] uses reverse distillation in both the training and prediction phases of the T–S network to promote the preferential acquisition of semantic knowledge to the student network. Inspired by this, we introduce skip connections and an attention module and use reverse distillation to improve the representation of feature map details since knowledge is then distilled to a shallow level.

Assume that the t-th block of the student network is D_{r}^t. DSKD first uses different sizes of convolution kernels (conv_1 and conv_2) for feature extraction in the previous teacher layer. We design different perceptual fields to obtain diverse feature results. The attention module inputs the results of conv_1, conv_2 from E_t, and D_{r}^t is obtained by connection with conv from D_{r}^{t+1}.d Using an attention mechanism helps D_{r} eliminate redundant information and learn the features related to the loss and the E's previous feature map. The feature computation in block D_{r}^t is shown in Eq. (2).

$$\begin{aligned} D_{\mathrm{r}}^t < &-\mathrm{conv}(D_{\mathrm{r}}^{t+1} \oplus \mathrm{att}(D_{\mathrm{r}}^{t+1}, (\mathrm{conv}_1(E^{t+1}) \\ &\oplus \mathrm{conv}_2(E^{t+1})))) \end{aligned} \tag{2}$$

Here, $\oplus$ denotes the connection, and the outermost convolution uses a $1 * 1$ convolution kernel. The stride for conv_1 and conv_2 is set to 2 to align with the input size of the next block. Key features from semantic knowledge D_{r}^{t+1} and teacher E^{t+1} are selected given to D_{r}^t to ensure that important anomaly patterns are captured. The loss function and anomaly score map generation are consistent with those used in forward distillation.

Anomaly score map fusion

For the treatment of anomaly score maps at different scales, methods such as [14] interpolate the results of small cosine similarity scores and use multiplication or addition to fuse them with large similarity scores, thus aggregating them into a single detection result. However, in our approach employing multiple student networks, a key issue surfaces: which anomaly score map from student nets is important. To solve this problem, we add synthetic noise in some regions of the image and regard this noise as anomaly mask M. The noise is added using an external data source A in the anomaly-free normal image I_n.

$$I_a = \beta(M \odot A) + (1 - \beta)(M \odot I_n) + (1 - M) \odot I_n\ , \tag{3}$$

where $\odot$ denotes the element-wise multiplication operations. β is opacity, which is regarded as data augmentation to increase the diversity of the training set, and it is randomly chosen between [0.15,1]. Such an injection of synthetic noise is also mentioned in DeSTSeg [49]. DeSTSeg uses a similar dual-student network for encoding and decoding, but it requires two residual blocks as a segmentation network; conversely, our work is more concerned with the aggregation of network results. We use an external dataset [10] for A, take

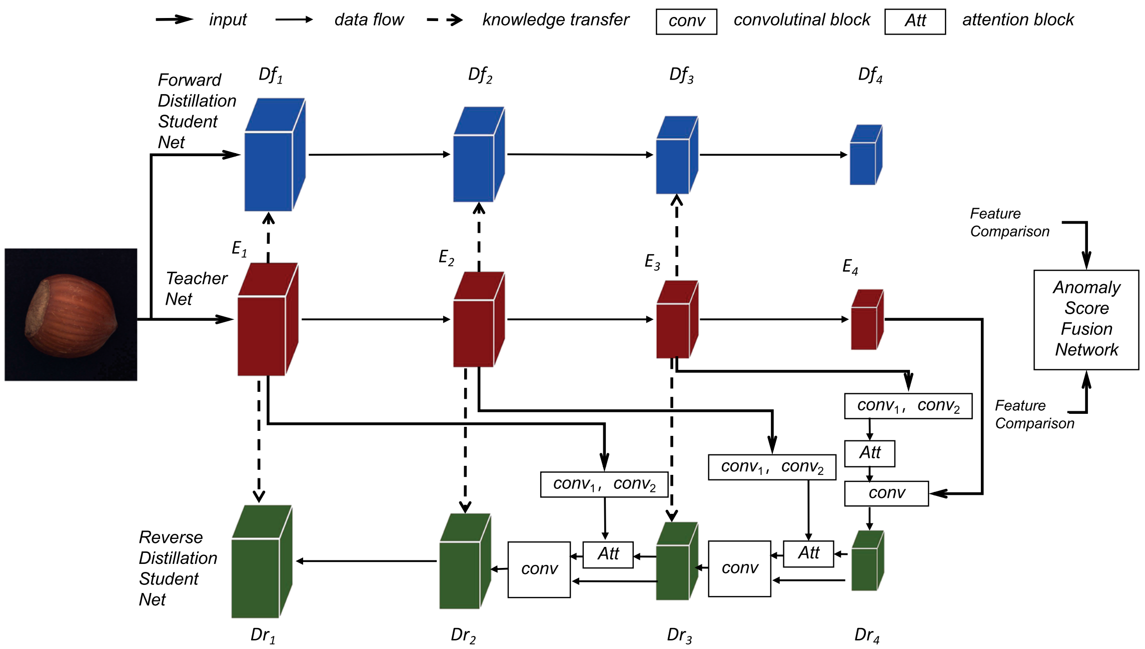

Fig. 1 Network architecture for DSKD. The realization represents the direction of data flow and the dotted line represents the anomaly score. The attention mechanism is denoted as att. Input images are fed through a teacher network and two student networks; the kernel sizes for conv_1 and conv_2 are $3 * 3$ and $7 * 7$, respectively. Feature comparison uses cosine similarity to compare each position in feature maps E_1 to E_3, providing a description of the inconsistency between the teacher network and the student network. Final result is computed by an anomaly score fusion network

the feature map of "Forward distillation student network"–"Reverse distillation student network" as input, and introduce dice loss [38] as the loss of the separating task. This loss is shown in Eq. (4).

$$\text{DL} = 1 - \frac{2 \sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2} \tag{4}$$

Here, p_i is the prediction associated with the ith pixel and g_i is its ground truth. The advantage of using dice loss is that we introduce a small percentage of synthetic masks that show an imbalanced distribution of semantic segmentation, which is in accordance with the small proportion of faults in one image. Dice loss focuses on mining foreground mask regions during training, and erroneous regions usually show range. As a region-based loss, dice loss also focuses on the correlation between the current and other pixel values. The anomaly map fusion network uses a $1 * 1$ convolutional kernel size to aggregate the anomaly score maps. Activation is performed using relu and sigmoid functions to ensure that the results are between 0 and 1. In order to preserve the overall image information, the amount of added noise is typically restricted. We use relu to suppress non-zero pixels from entering the sigmoid, thereby reducing the likelihood of the network failing to predict the background. We also note that using relu and sigmoid is a common approach for applying Dice loss [19, 42], while softplus cause the prediction value to exceed one, surpassing the mask value. For student's results, cosine similarity and interpolation are used to obtain anomaly score maps with consistent scales, and the final detection map is obtained using the trained fusion network. The process of anomaly score map fusion is shown in Fig. 2.

Experimental results and analysis

Implementation details

The dataset used is MVTec [6], which focuses on industrial inspections. It contains 15 categories of images divided into two types: object and texture. We use 3629 images for training (these are only normal samples), and 1725 images for testing (these are both normal and abnormal samples). All images are resized to $128 * 128$ and $256 * 256$, so that we obtain results corresponding to images of different sizes. To verify the effectiveness of our experiments on anomaly detection for different datasets, we also conducted one-class novelty detection in CIFAR10 [26] with the same settings as RD4AD, images are resize to 32*32, and evaluation with Sample AUROC.

For comparison with recently published results, the optimizer settings were kept consistent with RD4AD [14], and

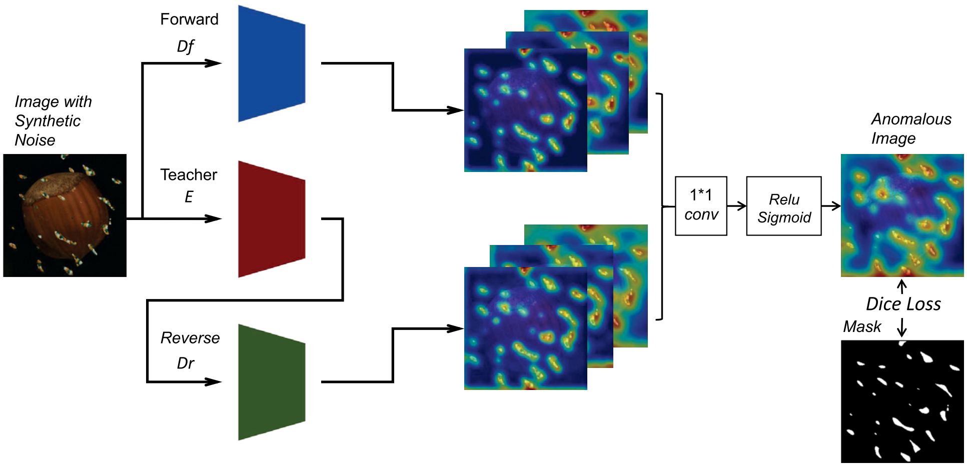

Fig. 2 Network architecture for anomaly score map fusion network. The anomaly-free image is given noise and is trained by its mask

we set an Adam optimizer with $\beta = (0.5, 0.999)$. The learning rate is set to 0.005, and we train for 200 epochs using an NVIDIA Tesla K40c and Intel(R) Xeon(R) E5-2680 CPU@2.80 GHz. Our method uses WideResNet50 as the teacher model with 68.8 M parameters. The forward-distilled student network has the same number of parameters as the teacher model, and the reverse-distilled student network has 108 M parameters. For comparison, the RD4AD model using the same teacher network structure has 67.2 M parameters for the BN layer and 24.9 M parameters for the student network. The number of parameters of our proposed method is about 1.58 times higher than that of RD4AD. This also implies that while DSKD can be trained simultaneously for both student networks, the training time is longer in the reverse network compared to the RD4AD. It should be noted that since the training phase only involves normal samples, the optimal number of epochs is less likely to be observed for comparison. For a fair and visual comparison, RD4AD is run on the same machine and with the same number of epochs. The result of the last epoch is used to construct the anomaly score map with a Gaussian filter applied for smoothness.

We evaluate the proposed method in terms of anomaly detection and anomaly localization. The receiver operating characteristic curve of samples (Sample AUROC) reflects the performance of the model in determining whether a sample is anomalous. We define the maximum value in the abnormal score map as Sample AUROC, which is also done in RD4AD. AUCROC for pixel (Pixel AUROC) and per-region overlap (PRO) reflect whether a pixel is judged as anomalous. Pixel AUROC calculates anomalies pixel by pixel, while PRO reduces the influence of overlapping anomalies. We compare our method with MKD [35], GT [18], GANomaly (GN) [2], Uninformed Student (US) [7], PSVDD [47], DAAD [20], MetaFormer (MF) [44], PaDiM [13], CutPastee [27], and RD4AD in the experiment of MVTec. The comparison methods of one-class novelty detection are LSA [1], OCGAN [34], HRN [21], DAAD and RD4AD. We analyze the advantages and limitations of our approach in the widely used anomaly detection dataset MVTec by evaluating semantic and pixel anomaly detection.

Anomaly detection

Table 1 shows the results of anomaly detection using Sample AUROC. Bold text in all tables in this paper indicates optimal values.

From Table 1, our proposed method achieves optimal average performance in texture images for different image sizes; DSKD follows behind RD4AD and PaDIM only for carpet and leather categories. Our method is valid for CIFAR10, which performs random additive synthetic noise on natural images to improve the metrics for one-class novelty detection and AUROC has reach 86.5%.

We also observed variations in the class distribution within the test set. When considering normal samples as negative samples and abnormal samples as positive samples, the positive-to-negative ratio ranges from 0.6 to 5.4. The transistor and pill categories are the smallest and largest categories, respectively. DSKD achieves the optimal Sample AUROC in both categories, thus showing that it is not much affected by sample imbalance in its anomaly detection performance. This finding also confirms that the anomaly score map fusion network consistently detects suspicious analog defects during training, and it effectively identifies normal samples that have not been previously encountered.

Limitations While DSKD achieves the average optimal performance for 128*128 images, the results for objects in 256*256 images are slightly worse than those obtained by RD4AD. We believe this is related to the added external data sources and the added locations. In texture images, the added locations are all targets to be detected while the objects contain part of the background. The added data sources for introducing synthetic noise are also textures and thus help textures improve the metrics. Moreover, the increase in image

Table 1 Sample AUROC for MVTec

Image size	128			256								
Category/ method	MKD	RD4AD	Ours	GN	US	PSVDD	DAAD	MF	PaDIM (ICPR21)	CutPast (CVPR21)	RD4AD (CVPR22)	Ours
Textures												
Carpet	79.3	**99.2**	98.7	69.9	91.6	92.9	86.6	94.0	**99.8**	93.9	98.9	99.2
Grid	78.0	95.7	**96.1**	70.8	81.0	94.6	95.7	85.9	96.7	**100.0**	**100.0**	**100.0**
Leather	95.1	**100.0**	99.2	84.2	88.2	90.9	86.2	99.2	**100.0**	**100.0**	**100.0**	**100.0**
Tile	91.6	99.4	**99.9**	79.4	99.1	97.8	88.2	99.0	98.1	94.6	99.3	**99.7**
Wood	94.3	98.8	**99.4**	83.4	97.7	96.5	98.2	99.2	99.2	99.1	99.2	**99.7**
Average	87.7	98.6	**98.7**	77.5	91.5	94.5	91.0	95.5	98.8	97.5	99.5	**99.7**
Objects												
Bottle	99.4	100.0	**100.0**	89.2	99.0	98.6	97.6	99.1	99.9	98.2	**100.0**	**100.0**
Cable	89.2	97.1	**99.7**	75.7	86.2	90.3	84.4	97.1	92.7	81.2	95.0	**99.4**
Capsule	80.5	**89.5**	84.3	73.2	86.1	76.7	76.7	87.5	91.3	98.2	**96.3**	91.6
Hazelnut	98.4	99.8	**100.0**	78.5	93.1	92.0	92.1	99.4	92.0	98.3	99.9	**100.0**
Metal Nut	73.6	99.2	**99.9**	70.0	82.0	94.0	75.8	96.2	98.7	99.9	**100.0**	**100.0**
Pill	82.7	93.3	**96.2**	74.3	87.9	86.1	90.0	90.1	93.3	94.9	96.6	**98.5**
Screw	83.3	91.1	**95.5**	74.6	54.9	81.3	**98.7**	97.5	85.8	88.7	97.0	95.4
Toothbrush	**92.2**	90.3	89.4	65.3	95.3	**100.0**	99.2	100.0	96.1	99.4	99.5	95.6
Transistor	85.6	99.5	**99.7**	79.2	81.8	91.5	87.6	94.4	97.4	96.1	96.7	**98.6**
Zipper	93.2	**94.3**	94.0	74.5	91.9	97.9	85.9	98.6	90.3	**99.9**	98.5	98.4
Average	87.8	95.4	**95.9**	75.5	85.8	90.8	88.8	96.0	93.8	95.5	**98.0**	97.8

size does not necessarily improve the Sample AUROC; this is evident for tile, screw, and transistor categories. Overall, DSKD is competitive in Sample AUROC, especially for textures.

Anomaly localization

Tables 2 and 3 present the Pixel AUROC and PRO results. The data in Tables 2 and 3 are similar in that DSKD achieves the best average performance in all texture categories; especially for 128*128 images, it performs better than RD4AD in all texture categories. For objects, although the results are not as significant as for textures, DSKD performs better than RD4AD in the *bottle* category. In terms of pixel AUROC, the optimal metrics of 10 and 6 out of 15 are achieved in the different sizes. Advantages in PRO come mainly with 128*128 images. By observing Tables 2 and 3, our proposed method has good competitiveness in texture categories.

Limitations DSKD has the third highest average result in PRO of all compared methods; moreover, its PRO is below 90 in *transistor* and *metalnut* categories. The location of defects in both categories is highly variable. For instance, *transistor* images have misplaced defects and *metalnut* images have flip defects. These types of defects usually cover the whole inspection area. DSKD is similar to [39] and it performs well in locating small anomalies in structurally simple samples but not so well in locating large anomalous regions in object categories. We believe that the modeling of noise is a factor. For objects with different-sized parts and possible defects, we use the same noise simulation strategy to help the model. Due to the lack of a priori knowledge of possible a priori faults, there is still room for improvement in this noise simulation strategy for samples with a large range of faults. To explore practical applications, we utilized the Gamma distribution [24] for automatic threshold screening and pixel-level anomaly detection in 128*128. The average recall is 15.8% higher than the accuracy, which is acceptable for recall-sensitive application.

Ablation study and visual analysis

Our proposed method introduces two key innovations: a novel network structure and a result fusion technique. The network structure includes a forward (Pos) and a reverse (Res) student network. Tables 4 and 5 present three metrics corresponding to the use of the two student networks for 128*128 images and 256*256 images, respectively. In addition, we manually set the weighting of the two student networks using a $1:3$ labeled manual fusion (MaFu) function. Feature fusion using synthetic noise is labeled synthetic fusion.

Table 2 Pixel AUROC for MVTec

Image size	128			256					
Category/method	MKD	RD4AD	Ours	SPADE	PaDiM	RIAD	CutPaste	RD4AD	Ours
Textures									
Carpet	95.6	98.0	**98.6**	97.5	**99.1**	96.3	98.3	98.9	**99.1**
Grid	91.8	97.0	**97.7**	93.7	97.3	98.8	97.5	98.9	**99.2**
Leather	98.1	99.0	**99.1**	97.6	97.6	99.4	**99.5**	99.4	99.3
Tile	82.8	92.6	**94.3**	87.4	94.1	89.1	90.5	95.6	**96.4**
Wood	84.8	92.2	**94.9**	88.5	94.9	85.8	95.5	95.4	**95.8**
Average	90.6	95.8	**96.9**	92.9	96.6	93.9	96.3	97.6	**98.0**
Objects									
Bottle	96.3	98.3	**98.7**	98.4	98.3	98.4	97.6	98.7	**98.9**
Cable	82.4	**97.8**	97.3	97.2	96.7	84.2	90.0	97.3	**98.1**
Capsule	95.9	96.3	**96.6**	**99.0**	98.5	92.8	97.4	98.6	98.8
HazeInut	94.6	**98.8**	98.7	**99.1**	98.2	96.1	97.3	98.9	99.0
Metal Nut	86.4	**96.6**	96.1	**98.1**	97.2	92.5	93.1	97.3	96.8
Pill	89.6	97.0	**97.7**	96.5	95.7	95.7	95.7	**98.3**	97.8
Screw	96.0	98.1	**98.8**	98.9	98.5	98.8	96.7	**99.6**	98.8
Toothbrush	96.1	**98.1**	97.6	97.9	98.8	98.9	98.1	**99.1**	98.8
Transistor	79.6	**97.7**	97.4	94.1	**97.5**	87.7	93.0	93.0	92.0
Zipper	93.9	96.8	**97.3**	96.5	98.5	97.8	**99.3**	98.2	97.7
Average	91.1	**97.6**	**97.6**	97.6	97.8	94.3	95.8	**97.9**	97.7

Table 3 POC for MVTec

Image size	128		256					
Category/method	RD4AD	Ours	US	MF	SPADE	PaDiM	RD4AD	Ours
Textures								
Carpet	95.5	**96.2**	87.9	87.8	94.7	96.2	96.9	**97.3**
Grid	92.2	**93.8**	95.2	86.5	86.7	94.6	96.8	**97.5**
Leather	**98.6**	**98.6**	94.5	95.9	97.2	97.8	**99.1**	99.0
Tile	85.1	**86.5**	**94.6**	88.1	75.9	86.0	90.3	92.3
Wood	81.9	**84.5**	91.1	84.8	87.4	91.1	90.7	**92.1**
Average	90.7	**91.9**	92.7	88.6	88.4	93.1	94.8	**95.6**
Objects								
Bottle	94.7	**95.4**	93.1	88.8	95.5	94.8	**96.6**	96.3
Cable	**90.5**	90.1	81.8	**93.7**	90.9	88.8	91.0	92.6
Capsule	**86.6**	85.9	**96.8**	87.9	93.7	93.5	95.8	94.4
Hazelnut	**89.6**	88.7	**96.5**	88.6	95.4	92.6	95.5	95.5
Metal Nut	**83.5**	82.2	94.2	86.9	**94.4**	85.6	92.4	88.2
Pill	90.0	**93.3**	96.1	93.0	94.6	92.7	**96.4**	95.5
Screw	93.8	**95.2**	94.2	95.4	96.0	94.4	**98.2**	95.2
Toothbrush	**86.5**	83.4	93.3	87.7	93.5	93.1	**94.2**	93.2
Transistor	**85.6**	84.7	66.6	**92.6**	87.4	84.5	79.4	79.6
Zipper	91.9	**92.8**	95.1	93.6	92.6	95.9	**95.6**	94.3
Average	**89.3**	89.2	90.8	90.8	93.4	91.6	**93.5**	92.5

Table 4 Metrics for forward distillation, reverse distillation, manual fusion and our method in 128 size

Evaluation metrics	Sample AUROC				Pixel AUROC				Pixel AUPRO			
Category/ method	Pos	Res	Manul fusion	Synthetic fusion	Pos	Res	Manul fusion	Synthetic fusion	Pos	Res	Manul fusion	Synthetic fusion
Textures												
Carpet	94.6	**99.2**	**99.2**	98.7	92.8	97.4	97.4	**98.6**	94.1	94.4	94.6	**96.2**
Grid	89.2	**96.5**	96.4	96.1	90.8	96.6	96.7	**97.7**	88.8	92.4	92.8	**93.8**
Leather	**100.0**	**100.0**	**100.0**	99.2	97.6	98.8	98.9	**99.1**	97.9	98.3	98.4	**98.6**
Tile	99.9	**100.0**	**100.0**	99.9	87.6	93.9	94.1	**94.3**	85.3	85.8	86.2	**86.5**
Wood	**99.4**	97.0	98.2	**99.4**	89.5	89.6	92.0	**94.9**	79.0	76.7	77.9	**84.5**
Average	92.2	98.5	**98.8**	98.7	91.2	95.3	95.8	**96.9**	82.9	89.5	90.0	**91.9**
Objects												
Bottle	99.6	**100.0**	**100.0**	**100.0**	97.4	97.7	97.8	**98.7**	93.9	92.6	92.8	**95.4**
Cable	97.3	99.4	99.5	**99.7**	91.7	**97.3**	**97.3**	**97.3**	85.4	89.9	90.0	**90.1**
Capsule	72.4	91.9	**93.2**	84.3	90.6	96.8	**98.0**	96.6	63.3	89.4	**92.8**	85.9
Hazelnut	99.1	99.9	99.9	**100.0**	96.3	98.5	98.6	**98.7**	86.5	88.1	88.1	**88.7**
Metal Nut	99.2	99.5	99.5	**99.9**	95.1	95.9	95.9	**96.1**	78.5	81.0	80.9	**82.2**
Pill	95.0	95.4	95.6	**96.2**	97.4	**97.7**	**97.7**	**97.7**	93.2	90.9	91.2	**93.3**
Screw	77.8	83.3	85.0	**95.5**	87.3	96.6	97.6	**98.8**	66.8	89.7	92.2	**95.2**
Toothbrush	88.3	91.4	**92.2**	89.4	96.4	97.6	**97.8**	97.6	80.0	**84.2**	84.0	83.4
Transistor	**100.0**	99.8	99.8	99.7	**98.0**	**98.0**	97.9	97.4	83.2	**85.6**	85.3	84.7
Zipper	94.9	**95.1**	**95.1**	94.0	90.4	95.8	95.8	**97.3**	79.7	90.2	90.2	**92.8**
Average	92.4	95.6	**96.0**	95.9	94.1	97.2	97.4	**97.6**	81.1	88.2	88.8	**89.2**

Table 5 Metrics for forward distillation, reverse distillation, manual fusion and our method in 256 size

Evaluation metrics	Sample AUROC				Pixel AUROC				Pixel AUPRO			
Category/ method	Pos	Res	Manul fusion	Synthetic fusion	Pos	Res	Manul fusion	Synthetic fusion	Pos	Res	Manul fusion	Synthetic fusion
Textures												
Carpet	**99.4**	98.0	99.2	99.2	97.3	98.6	98.7	**99.1**	95.3	94.8	96.5	**97.3**
Grid	99.7	99.5	99.5	**100.0**	97.8	96.4	98.9	**99.2**	95.2	96.6	96.6	**97.5**
Leather	**100.0**	**100.0**	**100.0**	**100.0**	98.2	**99.3**	99.2	**99.3**	98.6	98.8	98.9	**99.0**
Tile	**100.0**	**100.0**	**100.0**	99.7	93.8	95.4	95.5	**96.4**	90.9	90.9	91.2	**92.3**
Wood	99.2	98.7	99.5	**99.7**	91.3	94.5	94.2	**95.8**	84.6	88.2	89.5	**92.1**
Average	99.7	99.2	99.6	**99.7**	95.7	96.8	97.3	**98.0**	92.9	93.9	94.5	**95.6**
Objects												
Bottle	**100.0**	**100.0**	**100.0**	**100.0**	98.8	98.7	98.7	**98.9**	96.0	96.5	**96.6**	96.3
Cable	91.0	99.6	**99.7**	99.4	94.8	**98.3**	**98.3**	98.1	85.6	92.9	**93.0**	92.6
Capsule	44.2	98.3	**98.8**	91.6	75.4	98.7	**98.9**	98.8	61.5	95.7	**96.1**	94.4
Hazelnut	**100.0**	**100.0**	**100.0**	**100.0**	98.3	98.8	98.8	**99.0**	94.0	94.3	94.5	**95.5**
Metal Nut	88.8	**100.0**	99.1	**100.0**	93.9	**96.8**	93.1	**96.8**	72.7	86.8	**90.7**	88.2
Pill	95.9	97.8	98.0	**98.5**	98.2	**98.4**	98.3	97.8	90.8	95.5	**95.6**	95.5
Screw	83.5	**96.7**	92.9	95.4	86.0	**99.3**	98.3	98.8	55.1	**98.0**	93.0	95.2
Toothbrush	90.3	95.8	**96.1**	95.6	98.3	**98.9**	**98.9**	98.8	89.0	92.7	92.9	**93.2**
Transistor	**99.8**	99.1	99.1	98.6	**96.5**	93.6	93.1	92.0	79.8	80.8	**90.7**	79.6
Zipper	95.4	97.1	97.7	**98.4**	92.6	**98.0**	97.5	97.7	82.3	**94.8**	94.1	94.3
Average	88.9	**98.4**	98.1	97.8	93.3	**98.0**	97.4	97.7	80.7	92.8	**93.7**	92.5

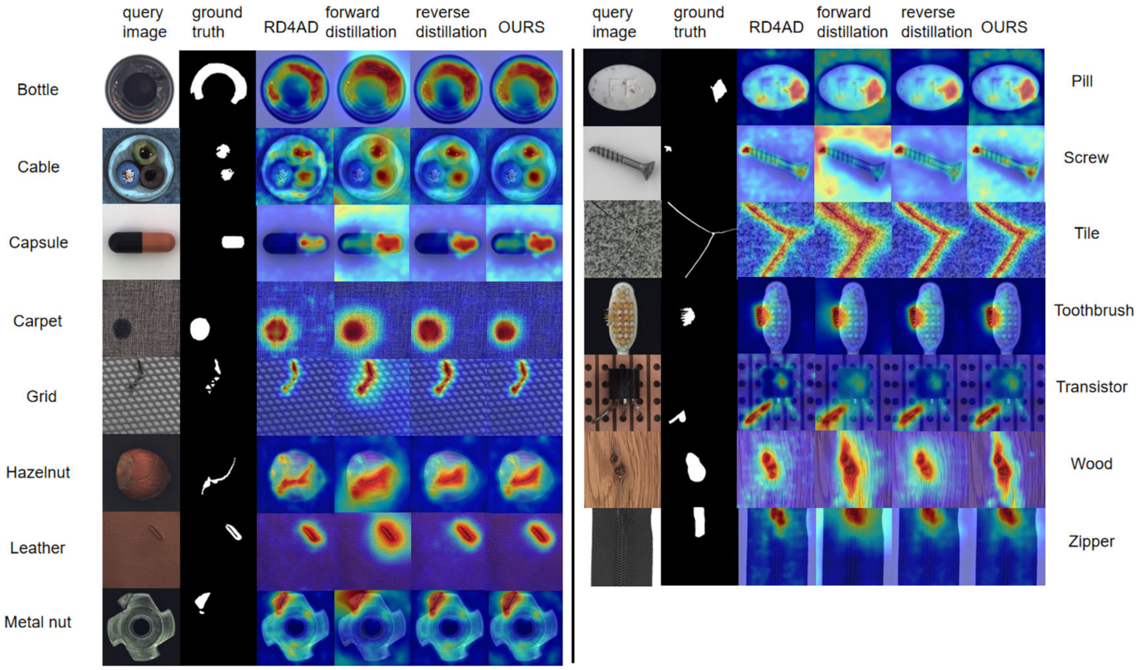

Fig. 3 Visualization results for MVTec, where the blue color represents pixels predicted as normal

Typically, the results of forward distillation are weaker than those of reverse distillation. However, there are some inconsistencies in the results for the *bottle* and *transistor* categories, which highlights the necessity of incorporating forward distillation in our DSKD approach. The metrics for reverse distillation are also better with our method, thus suggesting that our use of skip connections and an attention mechanism is effective. The pixel AUROC for manual fusion is inconsistent across the two image sizes. By contrast, DSKD with synthetic noise enhancement outperforms both the unidirectional distillation approaches and manual fusion in terms of pixel AUROC. It is shown that our proposed approach achieves good improvement in anomaly localization for dual-student networks.

In Fig. 3, we show the visualization results of forward distillation, reverse distillation, RD4AD, and our proposed method obtained on 15 categories with an image size of 256∗ 256.

Our results reveal that forward distillation tends to identify more defective regions, such as those in bottle and cable categories, compared with reverse distillation. This is also usually true compared with the regions identified by RD4AD. However, forward distillation also has some false defect regions, such as those seen in screw and metal nut categories. This phenomenon is not seen in our method. Our method also combines the advantages of forward and reverse distillation to obtain better defect prediction regions (see, for example, transistor and wood categories). Compared with RD4AD, our method has complete defect region prediction even in the case of there being a high probability of defects.

Discussion

Why two student models with different distillation patterns were applied Our dual-student training strategy uses normal samples and their training can be run in parallel. In Table 5, the single-model student network performs inconsistently across different categories. How to aggregate the two parts of detection is a novel problem. The forward and reverse distillation results in Fig. 3 are also inconsistent; in addition, forward distillation is usually more sensitive to anomalous regions than reverse distillation. Our model draws on the advantages of both types of distillation.

Why we used synthetic noise for model integration When employing multiple models, how to aggregate detection results is a key question. For anomaly detection, texture anomaly is an important component. We can combine the existing texture dataset by Eq. (3). The data can be quickly fused into the training process, and it only adds six 1*1 convolution parameters. This is acceptable for both training and testing. According to Tables 1, 2 and 3, such synthetic noise is helpful for detecting both texture-related anomalies and object-related anomalies. Our proposed synthetic noise enhances metrics and allows us to fuse results.

Conclusion

In this paper, we present a novel approach for anomaly detection called dual-student knowledge distillation. A reverse distillation network is constructed using skip connections and an attention module, which helps the reverse distillation network obtain detailed information and high-level representations. In addition, we construct a forward distillation network using a simple architecture. After combining the two distillation results, we introduce synthetic noise to help different abnormal score maps to adaptively assign weights. Experimental results demonstrate that our approach achieves state-of-the-art performance on the texture images from the MVTec dataset while also obtaining competitive scores on object images.

We believe that we can further improve object detection by further combining the results of object segmentation with noise to reduce the effect of noise on distinguishing an object from its background. Also, during the training annotation process, the region where the synthetic noise is added can be determined by manual annotation, while no changes are required during the inference process. Setting the probability threshold in abnormal detection is another area for further research since the high probability part of the model prediction region can cover most of the real defect regions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40747-024-01412-4.

Acknowledgements Kai Huang reports financial support was provided by Natural Science Foundation (3502Z202372018) of Xiamen, China. Kai Huang reports financial support was provided by Department of Education (JAT232012) of Fujian Province of China. Jian Mao reports financial support was provided by Natural Science Foundation (2023J01132745) of Fujian Province of China. Jian Mao reports financial support was provided by Xiamen Science and Technology Subsidy Project (2023CXY0318).

Data Availability Dataset is available at locations cited in the reference section.

Declarations

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Abati D, Porrello A, Calderara S et al (2019) Latent space autoregression for novelty detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). Long Beach, CA, USA, pp 481–490. https://doi.org/10.1109/cvpr.2019.00057
2. Akcay S, Atapour-Abarghouei A, Breckon TP (2019) Ganomaly: Semi-supervised anomaly detection via adversarial training. In: Computer Vision-ACCV 2018: 14th Asian conference on computer vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14. Springer, pp 622–637. https://doi.org/10.1007/978-3-030-20893-6_39
3. Akçay S, Atapour-Abarghouei A, Breckon TP (2019) Skip-ganomaly: skip connected and adversarially trained encoder–decoder anomaly detection. In: 2019 international joint conference on neural networks (IJCNN). IEEE, pp 1–8. https://doi.org/10.1109/ijcnn.2019.8851808
4. Alelaumi S, Wang H, Lu H et al (2020) A predictive abnormality detection model using ensemble learning in stencil printing process. IEEE Trans Compon Packag Manuf Technol 10(9):1560–1568. https://doi.org/10.1109/tcpmt.2020.3012501
5. Azad HK, Deepak A, Chakraborty C et al (2022) Improving query expansion using pseudo-relevant web knowledge for information retrieval. Pattern Recognit Lett 158:148–156. https://doi.org/10.1016/j.patrec.2022.04.013
6. Bergmann P, Fauser M, Sattlegger D et al (2019) Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). Long Beach, CA, USA, pp 9592–9600. https://doi.org/10.1109/cvpr.2019.00982
7. Bergmann P, Fauser M, Sattlegger D, et al (2020) Uninformed students: student-teacher anomaly detection with discriminative latent embeddings. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). Seattle, WA, USA, pp 4183–4192. https://doi.org/10.1109/cvpr42600.2020.00424
8. Cao Y, Wan Q, Shen W et al (2022) Informative knowledge distillation for image anomaly segmentation. Knowl Based Syst 248(108):846. https://doi.org/10.1016/j.knosys.2022.108846
9. Chen C, Li X, Huang K et al (2023) A convolutional autoencoder based fault detection method for metro railway turnout. CMES Comput Model Eng Sci. https://doi.org/10.32604/cmes.2023.024033
10. Cimpoi M, Maji S, Kokkinos I et al (2014) Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). Columbus, OH, USA, pp 3606–3613. https://doi.org/10.1109/cvpr.2014.461
11. Collin AS, De Vleeschouwer C (2021) Improved anomaly detection by training an autoencoder with skip connections on images corrupted with stain-shaped noise. In: 2020 25th international conference on pattern recognition (ICPR). IEEE, pp 7915–7922. https://doi.org/10.1109/icpr48806.2021.9412842
12. Das TK, Adepu S, Zhou J (2020) Anomaly detection in industrial control systems using logical analysis of data. Comput Secur 96(101):935. https://doi.org/10.1016/j.cose.2020.101935
13. Defard T, Setkov A, Loesch A, et al (2021) Padim: a patch distribution modeling framework for anomaly detection and localization. In: International conference on pattern recognition. Springer, pp 475–489. https://doi.org/10.1007/978-3-030-68799-1_35

14. Deng H, Li X (2022) Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). New Orleans, LA, USA, pp 9737–9746. https://doi.org/10.1109/cvpr52688.2022.00951
15. Deng J, Dong W, Socher R, et al (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 248–255. https://doi.org/10.1109/cvpr.2009.5206848
16. Dong H, Peng D (2018) Research on abnormal detection of modbustcp/ip protocol based on one-class svm. In: 2018 33rd Youth academic annual conference of chinese association of automation (YAC). IEEE, pp 398–403. https://doi.org/10.1109/yac.2018.8406407
17. Gogoi UR, Bhowmik MK, Bhattacharjee D et al (2018) Singular value based characterization and analysis of thermal patches for early breast abnormality detection. Australas Phys Eng Sci Med 41:861–879. https://doi.org/10.1007/s13246-018-0681-4
18. Golan I, El-Yaniv R (2018) Deep anomaly detection using geometric transformations. Adv Neural Inf Process Syst. https://doi.org/10.1145/3429309.3429326
19. Gros C, Lemay A, Cohen-Adad J (2021) Softseg: advantages of soft versus binary training for image segmentation. Med Image Anal 71(102):038. https://doi.org/10.1016/j.media.2021.102038
20. Hou J, Zhang Y, Zhong Q et al (2021) Divide-and-assemble: learning block-wise memory for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF international conference on computer vision (CVPR). Nashville, TN, USA, pp 8791–8800. https://doi.org/10.1109/iccv48922.2021.00867
21. Hu W, Wang M, Qin Q et al (2020) Hrn: a holistic approach to one class learning. Adv Neural Inf Process Syst 33:19111–19124. https://doi.org/10.1007/978-981-4021-75-3_9
22. Hu Y (2020) Design and implementation of abnormal behavior detection based on deep intelligent analysis algorithms in massive video surveillance. J Grid Comput 18:227–237. https://doi.org/10.1007/s10723-020-09506-2
23. Jiang Y, Cao Y, Shen W (2023) A masked reverse knowledge distillation method incorporating global and local information for image anomaly detection. Knowl Based Syst 280(110):982. https://doi.org/10.1016/j.knosys.2023.110982
24. Kawaguchi Y, Imoto K, Koizumi Y, et al (2021) Description and discussion on dcase 2021 challenge task 2: unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions. arXiv preprint arXiv:2106.04492
25. Kim D, Jeong D, Kim H et al (2022) Spatial contrastive learning for anomaly detection and localization. IEEE Access 10:17366–17376. https://doi.org/10.1109/access.2022.3149130
26. Krizhevsky A, Hinton G, et al (2009) Learning multiple layers of features from tiny images
27. Li CL, Sohn K, Yoon J et al (2021) Cutpaste: self-supervised learning for anomaly detection and localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). Nashville, TN, USA, pp 9664–9674. https://doi.org/10.1109/cvprR46437.2021.00954
28. Li N, Chang F, Liu C (2020) Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes. IEEE Trans Multimed 23:203–215. https://doi.org/10.1109/tmm.2020.2984093
29. Ma Y, Jiang X, Guan N et al (2023) Anomaly detection based on multi-teacher knowledge distillation. J Syst Archit 138(102):861. https://doi.org/10.1016/j.sysarc.2023.102861
30. Mohamed AB, Abouhawwash M, Mahapatra B, et al (2022) Responsible artificial intelligence based system to reduce greenhouse gas emissions in 6g networks
31. Naseer S, Saleem Y, Khalid S et al (2018) Enhanced network anomaly detection based on deep neural networks. IEEE Access 6:48231–48246. https://doi.org/10.1109/access.2018.2863036
32. Othman SB, Almalki FA, Chakraborty C et al (2022) Privacy-preserving aware data aggregation for iot-based healthcare with green computing technologies. Comput Electr Eng 101(108):025. https://doi.org/10.1016/j.compeleceng.2022.108025
33. Peng Z, Song X, Song S et al (2023) Hysteresis quantified control for switched reaction–diffusion systems and its application. Complex Intell Syst 9(6):7451–7460. https://doi.org/10.1007/s40747-023-01135-y
34. Perera P, Nallapati R, Xiang B (2019) Ocgan: one-class novelty detection using gans with constrained latent representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 2898–2906. https://doi.org/10.1109/cvpr.2019.00301
35. Salehi M, Sadjadi N, Baselizadeh S, et al (2021) Multiresolution knowledge distillation for anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 14902–14912. https://doi.org/10.1109/cvpr46437.2021.01466
36. Schlegl T, Seeböck P, Waldstein SM, et al (2017) Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International conference on information processing in medical imaging. Springer, pp 146–157. https://doi.org/10.1007/978-3-319-59050-9_12
37. Shen L, Tao H, Ni Y et al (2023) Improved yolov3 model with feature map cropping for multi-scale road object detection. Meas Sci Technol 34(4):045406. https://doi.org/10.1088/1361-6501/acb075
38. Sudre CH, Li W, Vercauteren T, et al (2017) Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep learning in medical image analysis and multimodal learning for clinical decision support: third international workshop, DLMIA 2017, and 7th international workshop, ML-CDS 2017, held in conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3. Springer, pp 240–248. https://doi.org/10.1007/978-3-319-67558-9_28
39. Tong G, Li Q, Song Y (2023) Two-stage reverse knowledge distillation incorporated and self-supervised masking strategy for industrial anomaly detection. Knowl Based Syst 273(110):611. https://doi.org/10.1016/j.knosys.2023.110611
40. Üzen H, Türkoğlu M, Yanikoglu B et al (2022) Swin-mfinet: swin transformer based multi-feature integration network for detection of pixel-level surface defects. Expert Syst Appl 209(118):269. https://doi.org/10.1016/j.eswa.2022.118269
41. Wang L, Yoon KJ (2021) Knowledge distillation and student-teacher learning for visual intelligence: a review and new outlooks. IEEE Trans Pattern Anal Mach Intell 44(6):3048–3068. https://doi.org/10.1109/tpami.2021.3055564
42. Wang L, Wang C, Sun Z et al (2020) An improved dice loss for pneumothorax segmentation by mining the information of negative areas. IEEE Access 8:167939–167949. https://doi.org/10.1109/access.2020.3020475
43. Wang R, Zhuang Z, Tao H et al (2023) Q-learning based fault estimation and fault tolerant iterative learning control for mimo systems. ISA Trans 142:123–135. https://doi.org/10.1016/j.isatra.2023.07.043
44. Wu JC, Chen DJ, Fuh CS et al (2021) Learning unsupervised metaformer for anomaly detection. In: Proceedings of the IEEE/CVF international conference on computer vision (CVPR). Nashville, TN, USA, pp 4369–4378. https://doi.org/10.1109/iccv48922.2021.00433
45. Xia X, Pan X, Li N et al (2022) Gan-based anomaly detection: a review. Neurocomputing 493:497–535. https://doi.org/10.1016/j.neucom.2021.12.093
46. Xu C, Gao W, Li T et al (2023) Teacher–student collaborative knowledge distillation for image classification. Appl Intell 53(2):1997–2009. https://doi.org/10.1007/s10489-022-03486-4

47. Yi J, Yoon S (2020) Patch svdd: patch-level svdd for anomaly detection and segmentation. In: Proceedings of the Asian conference on computer vision (ACCV). Kyoto, Japan. https://doi.org/10.1007/978-3-030-69544-6_23
48. Yin C, Zhang S, Wang J et al (2020) Anomaly detection based on convolutional recurrent autoencoder for iot time series. IEEE Trans Syst Man Cybern Syst 52(1):112–122. https://doi.org/10.1109/tsmc.2020.2968516
49. Zhang X, Li S, Li X et al (2023) Destseg: segmentation guided denoising student-teacher for anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). Vancouver, Canada, pp 3914–3923. https://doi.org/10.1109/cvpr52729.2023.00381
50. Zhou L, Zhang C, Wu M (2018) D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPR). Salt Lake City, UT, USA, pp 182–186. https://doi.org/10.1109/cvprw.2018.00034

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.