Multi-task Hybrid Knowledge Distillation for Unsupervised Anomaly Detection

Author

***Abstract*— Detecting both logical and structural anomalies in an unsupervised anomaly detection task is a significant challenge due to the inherent differences between the two types of anomalies. The use of two-branch knowledge distillation to deal with these two types of anomalies separately is a generalized approach. However, existing methods often design dual branches separately, which does not effectively utilize the shared information between these two branches. Also, due to the introduction of bottleneck layers, a large amount of detailed information is often lost during the reconstruction process, resulting in many false positives. To overcome these drawbacks, we structure the student network as a multi-task model to enhance its feature extraction capability, thereby improving its ability to distinguish between logical and structural anomalies, especially under the constraint of limited training data. Additionally, we incorporated a self-supervised distillation loss within the logical detection branch and trained the model using a hybrid distillation approach. By leveraging the differences in features between self-distillations to detect logical anomalies, we effectively minimized the false positives that often arise from image reconstruction blurring due to feature compression in the logical branch. We conducted experiments on three well-known anomaly detection datasets to demonstrate the effectiveness of our approach. In particular, on the challenging MVTec LOCO AD dataset, our method achieved impressive results with a pixel-level sPRO of 82.9% and an image-level AUROC of 91.0%. The source code are available at https://anonymous.4open.science/r/MHKD4AD-287C.**

***Index Terms*— Anomaly localization, Knowledge Distillation, Unsupervised learning**

I. Introduction

Unsupervised anomaly detection [1]–[5] refers to the effective identification and localization of anomalies in the inference phase by training the model only on anomaly-free images, which is an important and extensively researched area within computer vision. Anomalies in industrial detection tasks can be broadly classified into two categories: structural anomalies and logical anomalies, as illustrated in Fig. 1. Current methods [6], [7] in this area focus primarily on identifying structural anomalies in relatively simple scenarios. These defects typically include surface imperfections such as scratches, dents, or various forms of contamination. However, these methods have limitations in detecting logical anomalies under highly semantically complex conditions. Any sample that is missing or does not conform to this arrangement is an anomaly.

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment.

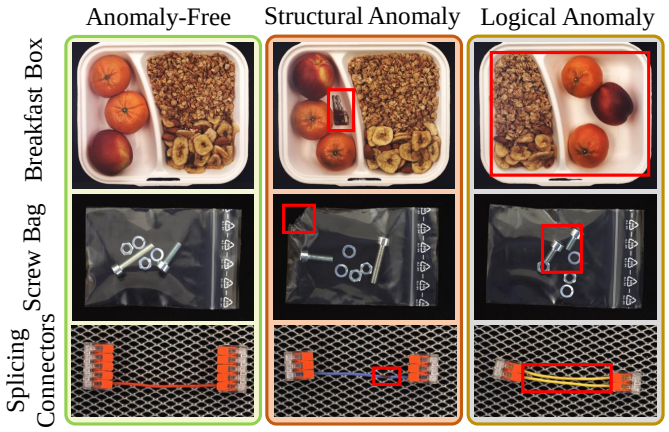
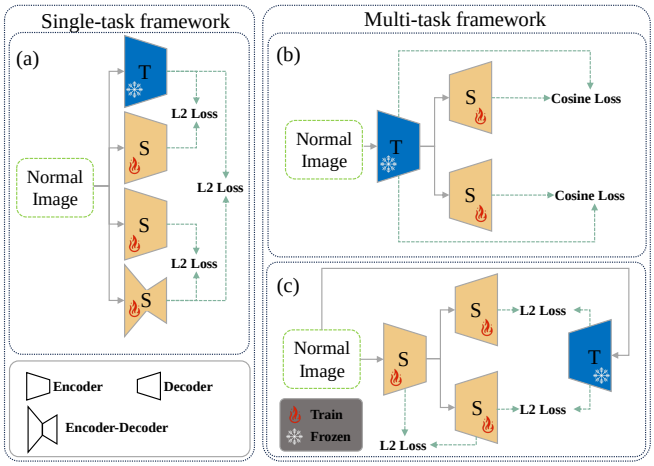


Fig. 1. Examples of different types of anomaly samples, including normal images (left), structural anomalies (middle), and logical anomalies (right). Structural anomalies introduce new local structures (such as a bunch of keys, broken plastic bags, and stained connecting wires), while logical anomalies violate the logical constraints of normal images (such as swapped breakfast boxes, incorrect numbers of long and short screws, and two connecting wires).

The complex and consistent arrangement within the image poses substantial challenges for existing anomaly detection algorithms, particularly in terms of accurately localizing and identifying logical anomalies. This complexity underscores the need for enhanced methods that can effectively interpret and process such intricate semantic structures.

Previous anomaly detection methods can be grouped into three main categories: feature representation-based [8]–[10], reconstruction-based [11]–[13] and distillation-based [14], [15]. During the training phase, feature representation-based methods use pre-trained networks to extract and store patch features of normal images. During testing, anomalies are detected by comparing the differences between the features of the input image and the stored features of normal images. These methods are particularly effective in dealing with structural anomalies. However, these methods have the drawback of losing global information between images, making them less effective in dealing with complex anomalies that violate logic. In addition, these methods store extensive patch feature data from normal images, with their efficiency and memory needs tied to the dataset size. Using an encoder-decoder scheme, reconstruction-based methods assume models trained on normal images struggle to accurately reconstruct anomalies. While effective for complex samples, this can lead to poor reconstruction and difficulties in detecting finer structural anomalies due to network generalization and feature compression.



(a) Framework of Efficient AD (b) Framework of DSKD (c) Framework of Ours

Fig. 2. Different strategies for anomaly detection.

Distillation-based approaches were first applied to anomaly detection in [6], [14], [16], combining the advantages of both feature representation-based and reconstruction-based methods. Compared to representation-based and reconstruction-based methods, these methodologies show outstanding effectiveness in detecting both logical and structural anomalies, which utilize a teacher network to guide the student network in extracting normal image features and identify anomalies by comparing the discrepancies between the teacher and student networks. Efficient AD [17] separates the detection of logical anomalies and structural anomalies into two distinct tasks and utilizes a multi-student network architecture to detect each type of anomaly independently. To detect structural anomalies, Efficient AD employs distillation encoder student networks and utilizes differences between teacher and student networks to facilitate detection. However, this design is almost ineffective in detecting logical anomalies. Since the structural branch lacks the ability to detect logical anomalies and the overall anomaly detection result still comes from the combination of the outputs of the two branches, the overall anomaly detection is not effective. DSKD [18] adopts a multi-task framework, where a teacher network guides the learning of two student networks to detect anomalies. For logical anomalies, the student network employs an autoencoder structure to memorize the features of normal images through the bottleneck structure. During the testing phase, it attempts to restore the anomalous image to a normal state for logical anomaly detection. However, this feature compression leads to the loss of detailed textures, resulting in blurred reconstructions and an increase in false positives in the detection results.

To address these limitations, we propose a novel approach for detecting and localizing anomalies by employing a multi-task hybrid knowledge distillation framework. As illustrated in Fig. 2, compared to the single-task and multi-task frameworks described in the preceding paragraph, we designed the student network as a multi-task model, employing a hard parameter-sharing strategy within the encoder. This design facilitates joint training across tasks for detecting logical and structural anomalies, enhancing the extraction of shared features. Our proposed method effectively addresses the limitations of existing methods in capturing both structural and logical anomalies and maximizes the use of limited training data, thereby improving overall detection accuracy. To strengthen logical anomaly detection, we introduced different levels of feature compression before the structural detection branch (LDecoder), bolstering its logical anomaly detection capabilities. Additionally, the bottleneck layer in the GDecoder often complicates the reconstruction task and may lead to misclassification. To address this issue, we utilized the final output of the logical detection branch as supervisory information for the output of the shallow encoder in the student network, resulting in the SEncoder1 generation of a smoothed representation of the teacher network's feature output. This strategy effectively aligns with the principles of a self-distillation mechanism. This approach detects logical anomalies by leveraging the differences between the logical detection branch and the student encoder, which strategy enhances anomaly detection performance without introducing additional modules or increasing computational overhead.

Our contributions can be summarized as follows:

- We propose an efficient teacher-student network structure that utilizes multi-task student networks for unsupervised anomaly detection tasks in complex situations.
- We introduce a self-distillation loss that enhances the student network's feature extraction capabilities, leading to a significant improvement in the performance of logical anomaly detection.
- To validate the effectiveness of the method, we conducted experiments on three publicly anomaly detection datasets, on the MVTec LOCO AD dataset we achieved state-of-the-art performance (pixel-sPRO 82.9%, image-AUROC 91.0%), and on the MVTec AD and ViaA datasets we achieved results with competitive results.

II. Related Work

We briefly review recent research in unsupervised anomaly detection, which can be categorized into three approaches: feature representation-based, reconstruction-based, and distillation-based.

Feature representation-based methods. Reiss et al. [9] propose a method that utilizes a pre-trained network to extract the feature representation of an image and model its distribution. During inference, anomalies are detected by analyzing the divergence between the sample's feature representation and the distribution of normal samples. However, this approach heavily relies on the quality of the pre-trained network and may struggle with subtle anomalies that have minimal deviation from the learned normal distribution. Additionally, the method does not explicitly address the challenge of feature compression, which can impact detection accuracy in cases with complex or diverse anomaly types.PaDiM [19] improves upon this by storing the Multivariate Gaussian distributions of the feature representations and calculating the Mahalanobis distance between test samples and the stored normal distributions for anomaly detection. While this approach benefits from using a statistical distance metric, its reliance on the Gaussian assumption limits its effectiveness when dealing with non-Gaussian data distributions. Moreover, the

storage of Multivariate Gaussian parameters for each feature map can lead to increased memory usage, making it less suitable for large-scale applications. Patch Core [8] introduces a patch description method that more effectively characterizes the extracted features. A key innovation of Patch Core is the utilization of a greedy algorithm to optimize the normal feature library. Instead of maintaining a large, exhaustive set of normal features, which can be computationally expensive and inefficient, Patch Core selectively reduces the size of this library. The greedy algorithm systematically identifies and retains the most representative normal features, thereby streamlining the feature library without sacrificing descriptive power. Consequently, Patch Core achieves a delicate balance between maintaining high detection accuracy and ensuring operational efficiency, addressing some of the critical challenges in the field of unsupervised anomaly detection. However, Patch Core may still encounter challenges in detecting anomalies that appear in low-frequency or sparsely sampled regions of the feature space, as the greedy selection process could potentially exclude subtle but critical features. Additionally, the method requires careful tuning of the number of retained features to avoid compromising detection performance. Overall, these methods provide important advancements in unsupervised anomaly detection, yet they exhibit limitations in handling complex data distributions, feature compression, and balancing detection accuracy with computational resources.

The reconstruction-based methods. An et al. [12], [13] are trained to minimize the reconstruction error of normal images. These methods rely on accurately reconstructing normal images while failing to do so for abnormal images during the testing phase. Common approaches include Generative Adversarial Networks (GANs) [20] and Variational Auto-Encoders (VAEs) [21], both of which are based on the principle of reconstructing normal patterns and identifying deviations. However, these reconstruction-based methods face significant challenges, particularly in scenarios where the anomalies share similar characteristics with the normal data. In such cases, the models may inadvertently reconstruct anomalous features, leading to a failure in detecting anomalies effectively. SCADN [22] extends this concept through inpainting frameworks that train models on data with masked normal regions, enabling them to leverage contextual information to reconstruct unseen areas. While this approach aids in detecting anomalies by evaluating the quality of the reconstructed regions, there remains an inherent limitation due to the strong generalization ability of deep learning models. These models are often capable of reproducing features that closely resemble the training data, even if they include subtle anomalies that were not explicitly present during training. This can lead to the unintended reconstruction of anomalous features, diminishing the visibility of anomalies and reducing the accuracy and robustness of the detection process.

Distillation-based methods. Bergmann et al. [14] use the difference between teacher and student networks to detect anomalies, leveraging the discrepancy between the networks' outputs to identify deviations from normal patterns. However, this approach may struggle to detect subtle or contextually complex anomalies where the differences are not pronounced enough to trigger detection. RD [15] improves upon this concept by introducing reverse distillation networks with an encoder-decoder structure, aiming to enhance the performance of anomaly detection. Despite the improvements, the encoder-decoder design still faces challenges related to feature compression and information loss, which can reduce the accuracy when detecting anomalies with fine details. To address the issue of reconstruction detail loss, THFR [13] uses a Template-based approach to help the image recover details by referencing a normal image as a kind of template. While this method can improve the quality of reconstructed abnormal images by guiding the recovery process, it also introduces a significant drawback. The differences between the normal template and the input image can negatively impact the reconstruction quality for normal images, often resulting in incomplete or inaccurate reconstruction. During testing, the Template may aid in reconstructing an abnormal image to appear normal, but for normal images, it can be counterproductive, leading to a failure in preserving fine details and, consequently, a decrease in detection accuracy.

Summary: Our model integrates a novel hybrid distillation strategy that combines self-distillation and knowledge distillation, effectively reducing false positives caused by low reconstruction quality in the logical branch. Specifically, self-distillation supervises the shallow features of the student network with its own outputs, identifying logical anomalies by measuring inconsistencies between them. In contrast to conventional approaches that rely solely on knowledge distillation, our model employs a multi-task knowledge distillation framework. This framework overcomes critical limitations in unsupervised anomaly detection, which is inadequate handling of diverse anomaly types. By addressing these challenges, our method significantly enhances the robustness and adaptability of anomaly detection systems.

III. Proposed Method

In this paper, we propose a multi-task hybrid knowledge distillation model for anomaly detection, the architecture of which is detailed in Fig. 3. In the training phase, the teacher network F_t guided the learning of both the logical and structural decoder branches within the student network F_s. Additionally, the logical decoder branch, GDecoder, supervised the learning of the student network's low-level encoder, SEncoder1. During the testing phase, structural anomaly maps $\mathcal{A}_s$ and logical anomaly maps $\mathcal{A}_g$ are generated by comparing features from the teacher with those from the structural (LDecoder) and logical (GDecoder/SEncoder1) components, respectively. Finally, we were able to combine the two anomaly scores to obtain the final anomaly score $\mathcal{A}$.

A. Teacher Network

The architecture of our teacher network F_t is composed of six convolutional layers coupled with two average pooling layers. This configuration is analogous to that utilized in the Efficient AD [17], where each neuron in the output layer of F_t encompasses a receptive field spanning 33×33 pixels. This design prevents anomalies in one region from affecting distant, unrelated areas, thereby enhancing localization accuracy

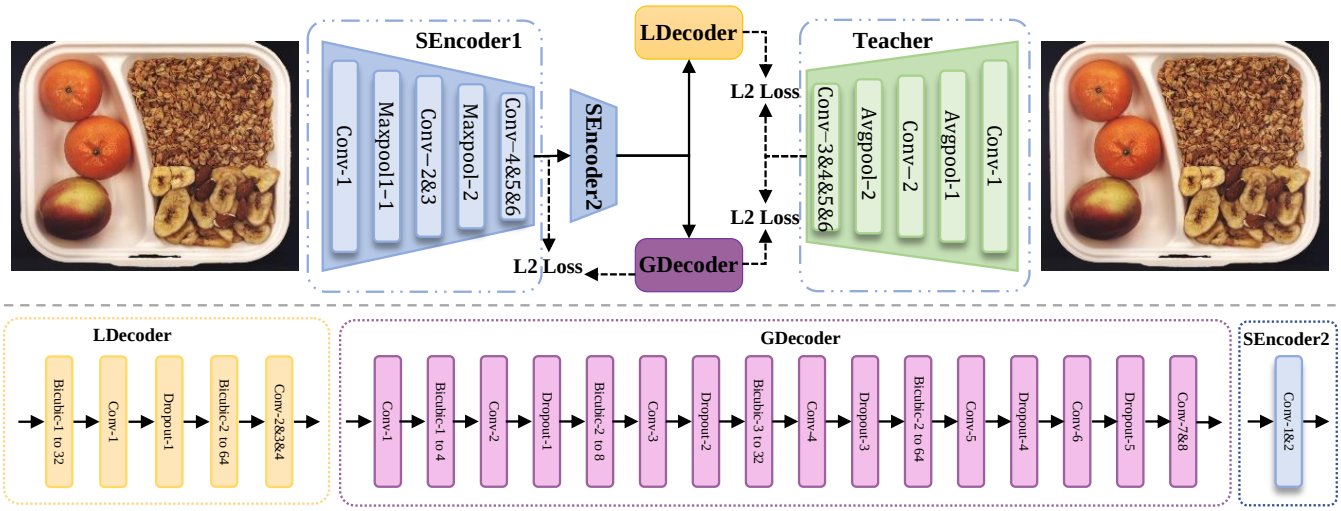


Fig. 3. Overview architecture of proposed Multi-task Hybrid Knowledge Distillation Framework.

in anomaly detection tasks. F_t is distillated from a pre-trained WideResNet-101 model [23], leveraging the extensive ImageNet dataset. During the distillation process, we utilize the mean square error (MSE) as the loss function to guide the learning of F_t.

To enhance the efficacy of our model, we adopt the feature post-processing strategy from PatchCore [8], and then, each feature vector is mapped into a compressed 384-dimensional space, where the feature representation is optimized. We obtain the features $f_t \in \mathbb{R}^{h\times w\times c}$ by feeding the training images x into the F_t, defined as:

$$f_t = F_t(x). \tag{1}$$

B. Student Network

The student network F_s comprises four components: SEncoder1, SEncoder2, LDecoder, and GDecoder. SEncoder1 includes six convolutional layers and two maximal pooling layers. The configuration of convolutional and max-pooling layers in SEncoder1 differs substantially from that of the teacher network, with variations in both layer arrangement and channel count. Such a deliberately engineered asymmetry in the architectural design is strategically implemented to enhance the network's sensitivity to anomalous images. This enhancement significantly improves the overall accuracy of anomaly detection, as elaborated in [24]. We extract the shallow sub-features f_l of the training image x by adopting the SEncoder1 network. Subsequently, SEncoder2, a simplified encoder with two convolutional layers, processes this sub-feature f_l to generate an enhanced feature map f_h, which can be written as:

$$\begin{aligned} f_l &= \text{SEncoder1}(x), \\ f_h &= \text{SEncoder2}(f_l). \end{aligned} \tag{2}$$

To address the distinct challenges of structural and logical anomaly detection, we designed a two-path decoder framework. The LDecoder branch is mainly employed to detect structural anomalies, while the GDecoder branch detects logical anomalies by analyzing global image features to identify deviations from standard composition and coherence. The structure of the Gdecoder network consists of convolutional layers, upsampling layers, and Dropout layers. The feature map f_h is fed into the bottleneck layer Conv-1 which uses a convolutional kernel of size 16, to compress the input into a one-dimensional global feature representation $f_{g'} \in \mathbb{R}^{1\times 1\times 64}$ when the input image size is 256. The fundamental difference between the GDecoder and LDecoder branches lies in feature compression strategy. This strategy compresses image features into a one-dimensional vector, effectively preventing the reconstruction of anomalies during the testing phase. By doing so, it forces the network to memorize only the prominent features of normal images during the training process. After the initial extraction, the global feature $f_{g'}$ is subjected to a meticulous restoration process facilitated by the GDecoder. This restorative procedure aims to transform $f_{g'}$ into the features f_g, aligning them with the spatial resolution of the feature map f_t. This process is characterized as follows:

$$f_g = \text{GDecoder}(f_h). \tag{3}$$

A dropout layer is also incorporated into this architecture, strategically designed to mitigate the risk of network overfitting.

The computation of the reconstruction loss for the GDecoder is formulated as follows:

$$\mathcal{L}_G = (hwc)^{-1}\sum_c \|f_t - f_g\|_F^2. \tag{4}$$

LDecoder, comprising upsampling, convolutional, and Dropout layers, takes the output feature f_h from SEncoder2 as input to generate structure feature f_s, as defined by the following formula:

$$f_s = \text{Lecoder}(f_h). \tag{5}$$

The structural loss function $\mathcal{L}_L$ is formulated as follows:

$$\mathcal{L}_L = (hwc)^{-1}\sum_c \|f_t - f_s\|_F^2. \tag{6}$$

We design a self-supervised distillation loss $\mathcal{L}_{SG}$ to enhance feature consistency across different components, defined as follows:

$$\mathcal{L}_{SG} = (hwc)^{-1} \sum_{c} \|f_g - f_l\|_F^2. \tag{7}$$

The overall loss of the model is:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_L + \beta\mathcal{L}_G + \gamma\mathcal{L}_{SG}, \tag{8}$$

where α , β, and γ are used to adjust the relative contributions of the three loss terms.

C. Anomaly map

In the testing phase, we use the difference between SEncoder1 and Gdecoder as logical anomaly score $\mathcal{A}_g$.

The structural anomaly scores $\mathcal{A}_s$ are computed by comparing the output feature f_s of the LDecoder with the feature f_t from the teacher network F_t, indicating the model's capability to detect structural anomalies. After obtaining both the structural anomaly scores $\mathcal{A}_s$ and the logical anomaly scores $\mathcal{A}_g$, we normalize them to ensure that they are on similar scales. This normalization is critical because the disparity in scale could cause noise from one graph to obscure accurate detection in the other when combined. Following the approach used in Efficient AD, we compute the set of all pixel anomaly scores in the validation image for both anomaly maps. Then, two p-quantiles are computed separately for each set, and a linear transformation is performed for each. The structural and logical anomaly scores are then normalized by their respective linear transformations. Finally, the normalized scores are interpolated to the original image size and then summed to give the overall anomaly score, which is denoted as $\mathcal{A}$. The process can be written as follows:

$$\mathcal{A} = \varphi(\mathcal{A}_s) + \varphi(\mathcal{A}_g), \tag{9}$$

where φ represents the linear interpolation operation.

IV. Experiments

A. Experimental setup

To validate the effectiveness of our proposed method, the empirical evaluations on the MVTec LOCO AD, MVTec AD, and VisA datasets are performed.

MVTec LOCO AD dataset [28], specifically tailored for unsupervised anomaly detection, was provided by MVTec in Germany. MVTec LOCO AD dataset is designed to challenge and benchmark state-of-the-art methods in the field. Notably, the MVTec Logical Constraints Anomaly Detection dataset is currently one of the most challenging datasets available. It simulates real-world scenarios in industrial inspection, with data collected and produced directly from production lines. Structural anomalies, such as scratches, dents, and contamination, are common occurrences in industrial settings. Logical anomalies, on the other hand, involve misplacements of objects or the appearance of allowed objects in invalid locations. This dataset encompasses five authentic sub-datasets, featuring a training set with 1772 normal images for training, 304 normal images for validation, and 1568 images for testing.

VisA dataset [32]is comprised of 12 subsets, each corresponding to different objects, as illustrated in Fig. 5. In total, there are 10,821 images, including 9,621 normal samples and 1,200 anomalous samples. Among these subsets, four represent various types of printed circuit boards (PCBs), which feature relatively complex structures incorporating components like transistors, capacitors, and chips. Additionally, the dataset includes four subsets—Capsules, Candles, Macaroni1, and Macaroni2—where multiple instances within a single view vary significantly in their locations and poses. Another four subsets consist of Cashew, Chewing Gum, Fryum, and Pipe Fryum, where the objects are more roughly aligned. The anomalous images capture a range of flaws, encompassing surface defects such as scratches, dents, color spots, and cracks, as well as structural defects like misplaced or missing parts. Each defect type is represented by 5 to 20 images, and it is common for a single image to display multiple defects.

MVTec AD dataset [33] is one of the most challenging anomaly detection datasets mainly dominated by structural anomalies. These anomalies may arise due to object damage, defects, deformations, or other structural issues. It consists of 15 real-world sub-datasets including 5 classes of textures and 10 classes of objects.

Retinal OCT Dataset [34], from the spectralis OCT system (Heidelberg Engineering, Germany), serves as a vital resource for evaluating anomaly detection in retinal images. This dataset comprises four distinct categories: choroidal neovascularization (CNV), diabetic macular edema (DME), Drusen, and a normal control group. For a balanced and fair comparison, the dataset is meticulously divided into training and test sets by the publisher. The training set consists of a substantial 26,315 normal images. The test set is composed of 1,000 images—250 normal and 750 abnormal—encompassing the three mentioned disease categories (CNV, DME, and Drusen). Utilizing the normal images from the original training set, our model is trained to recognize the standard retinal features. Subsequently, we evaluate performance on the complete test set, enabling a comprehensive assessment of our model's ability to accurately identify anomalies.

Training details. We implement our method based on the PyTorch framework and train it from scratch using a machine with one NVIDIA GeForce RTX 3090 GPU. All images are resized to a resolution of 256×256. The one-model-per-category setting from previous studies is followed. For dataset segmentation, we use the pre-defined splits for training, validation, and testing provided with the publicly available datasets, ensuring consistency across all experiments and comparability with prior work. Each student is trained for 80,000 epochs with a batch size of 1. The Adam optimizer is used with a learning rate of 10^{-4} and weight decay of 10^{-5}. When 76,000 rounds have passed, the learning rate is reduced to 10^{-5}. During the training phase, the dropout ratio is set at 0.2. The anomaly score map $\mathcal{A}$ is resized back to the original image resolution using bilinear interpolation.

To assess the effectiveness of our proposed approach, we include a variety of classical and state-of-the-art (SOTA) methods for comparison:

- AE [12], VAE [21] and f-AnoGAN [20] are well-known reconstruction-based techniques. It employs encoder-decoder architecture, utilizing image reconstruction errors

TABLE I
QUANTITATIVE RESULTS ON MVTEC LOCO AD DATASET FOR ANOMALY LOCALIZATION, AS MEASURED ON PIXEL-SPRO. THE BEST RESULTS ARE MARKED IN BOLD.

Method	Breakfast Box	Screw Bag	Pushpins	Splicing Connectors	Juice Bottle	Mean
VM [25]	0.168	0.253	0.254	0.125	0.325	0.225
f-AnoGAN [20]	0.223	0.348	0.336	0.195	0.569	0.334
MNAD [26]	0.080	0.344	0.357	0.442	0.472	0.339
AE [12]	0.189	0.289	0.327	0.479	0.605	0.378
VAE [21]	0.165	0.302	0.311	0.496	0.636	0.382
SPADE [27]	0.372	0.331	0.234	0.516	0.804	0.451
S–T [14]	0.496	0.602	0.523	0.698	0.811	0.626
RD [15]	0.560	0.535	0.577	0.701	0.837	0.642
PaDiM [19]	0.509	0.461	0.295	0.467	0.779	0.502
Patch Core [8]	0.451	0.562	0.423	0.598	0.694	0.546
GCAD [28]	0.502	0.558	0.739	0.798	0.910	0.701
SimpleNet [29]	-	-	-	-	-	0.363
FastFlow [30]	-	-	-	-	-	0.568
DRAEM [31]	0.499	0.490	0.493	0.673	0.800	0.591
Efficient AD [17]	-	-	-	-	-	0.798
DSKD [18]	0.568	0.627	0.825	0.767	0.865	0.730
THFR [13]	0.583	0.615	0.763	0.848	0.896	0.741
Ours	**0.693**	**0.693**	**0.891**	**0.902**	**0.967**	**0.829**

or feature reconstruction errors to identify anomalies.

- SPADE [27], Patch Core [8] and PaDiM [19] record the feature representations of normal samples during training and detect anomalies by comparing test sample representations against these stored features.
- VM [25] and MNAD [26] are traditional methods of anomaly detection.
- S-T [14], RD [15], GCAD [28], DSKD [18], THFR [13], and Efficient AD [17] transfer knowledge of normal patterns from pre-trained teacher network to a simpler student network, enhancing anomaly detection performance.
- DRAEM [31] and SimpleNet [29] introduce artificially generated defective images into the training process, enabling models to learn to differentiate between these synthetic anomalies and normal examples.
- FastFlow [30] relies on flow-based models to estimate the density of normal data. Normal examples are expected to exhibit high likelihood under the learned distribution, while anomalies fall outside this range.

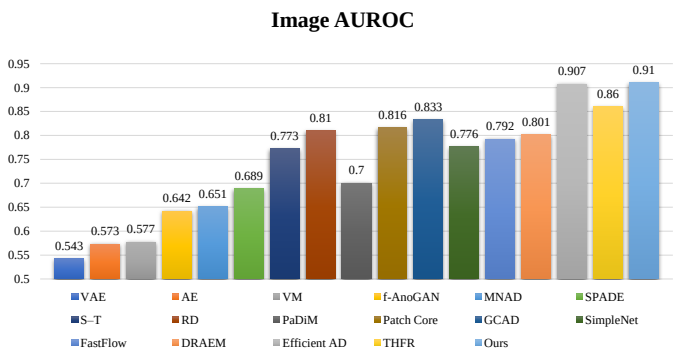


Fig. 4. Quantitative results on MVTec LOCO AD dataset for anomaly detection, as measured on the average image-AUROC.

Evaluation metrics.

To evaluate no-threshold image-level anomaly detection, we use the AUROC as the main metric, which measures an algorithm's ability to distinguish between normal and anomalous samples. To evaluate anomaly localization, AUROC is a suitable metric as it can measure algorithm performance in detecting structural anomalies. However, for logical anomalies such as missing objects, annotating and segmenting each pixel is difficult. The saturated per-region overlap (sPRO) metric [28] is used. This is an extended version of the PRO metric used to assess anomaly localization performance.

B. Anomaly detection and localization

Results on MVTec LOCO AD

We compared our proposed method with the current state-of-the-art techniques, and the quantitative results are summarised in Table I. Experimental data for all comparison methods were derived from the results reported in the respective papers. As shown in Table I, it is clear that distillation-based methods (e.g., GCAD [28], THFR [13], and Efficient AD [17]) outperform others in detecting both structural and logical anomalies. In comparison to other representation- and reconstruction-based methods, our proposed approach achieves the best performance in terms of pixel-sPRO and image-level AUROC. Specifically, the pixel-sPRO surpasses the state-of-the-art by an average of 3.1%, reaching 82.9%. As illustrated in Fig. 4, with an image AUROC of 0.91, our method demonstrates a high level of reliability in distinguishing between normal and anomalous images. The clear advantage over Efficient AD, despite its already remarkable performance, underscores the superior robustness and precision of our approach.

Our method also achieved clear advantages in challenging scenarios, such as breakfast boxes, screw bags, splicing connectors, and pushpins, where complex contextual logic constraints present considerable difficulties for other techniques. In the

TABLE II
QUANTITATIVE RESULTS ON VISA AND MVTEC AD DATASET FOR ANOMALY LOCALIZATION, AS MEASURED ON IMAGE-AUROC/PIXEL-AUROC.

Category		PatchCore [8]	RD [15]	RD++ [35]	SimpleNet [31]	DRAEM [29]	FastFLow [30]	DMAD [36]	Ours
ViaA	candles	98.6/99.5	92.2/97.9	96.4/98.6	98.7/97.7	94.4/97.3	92.8/94.9	92.7/98.1	97.9/98.7
	Capsules	81.6/99.5	90.1/89.5	92.1/99.4	89.9/99.0	76.3/99.1	71.2/75.3	88.0/99.2	93.2/99.7
	Cashew	97.3/98.9	99.6/95.8	97.8/95.8	97.5/98.8	90.7/88.2	91.0/91.4	95.0/95.3	96.9/97.9
	Chewing gum	99.1/99.1	99.7/99.0	96.4/99.4	99.8/98.3	94.2/97.1	91.4/98.6	97.4/97.9	98.9/98.8
	Fryum	96.2/93.8	96.6/94.3	95.8/96.5	98.1/91.1	97.4/92.7	88.6/97.3	98.0/97.0	96.2/97.0
	Macaroni1	97.5/99.8	98.4/97.7	94.0/99.7	99.4/99.6	95.0/99.7	98.3/97.3	94.3/99.7	96.8/99.9
	Macaroni2	78.1/99.1	97.6/87.7	88.0/99.7	82.4/98.9	96.2/99.9	86.3/89.2	90.4/99.7	93.0/99.7
	Pcb1	98.5/99.9	97.6/75.0	97.0/99.7	99.0/99.6	54.8/90.5	77.4/75.2	95.8/99.8	99.5/99.9
	Pcb2	97.3/99.0	91.1/64.8	97.2/99.0	99.1/98.3	77.8/90.5	61.9/67.3	96.9/99.0	99.3/99.2
	Pcb3	97.9/99.2	95.5/95.5	96.8/99.2	98.5/99.2	94.5/98.6	74.3/94.8	98.3/99.3	98.1/99.5
	Pcb4	99.6/98.6	96.5/92.8	99.8/98.6	99.6/93.9	93.4/88.0	80.9/89.9	99.7/98.8	99.6/98.9
	Pipe_fryum	99.8/99.1	97.0/92.0	99.6/99.1	99.7/98.9	99.4/90.9	72.0/87.3	99.0/99.3	100/99.4
	Mean	95.1/98.8	96.0/90.1	95.9/98.7	96.8/97.8	88.7/94.4	82.2/88.2	95.5/98.6	97.5/99.1
MVTec	Carpet	98.7/99.0	98.9/98.8	100/99.2	99.0/98.4	97.0/95.5	87.5/98.0	100/99.1	99.2/98.1
	Grid	98.2/98.7	100/97.0	100/99.3	97.8/98.9	99.9/99.7	88.1/93.5	100/99.2	100/98.3
	Leather	100/99.3	100/98.6	100/99.4	99.3/98.0	100/98.6	96.1/93.0	100/99.5	99.9/99.1
	Tile	98.7/95.6	99.3/98.9	99.7/96.6	99.7/96.6	99.6/99.2	63.2/96.1	100/96.0	100/97.8
	Wood	99.2/95.0	99.2/99.3	99.3/95.8	99.9/93.2	99.1/96.4	90.8/95.9	100/95.5	99.7/95.8
	Bottle	100/98.6	100/99.0	100/98.8	100/94.0	99.2/99.1	100/92.7	100/98.9	100/99.1
	Cable	99.5/98.4	95.0/99.4	99.2/98.4	99.9/98.7	91.8/94.7	91.2/97.5	99.1/98.1	99.3/98.8
	Capsule	98.1/98.8	96.3/97.3	99.0/98.8	100/97.4	98.5/94.3	99.2/97.4	98.9/98.3	98.3/99.2
	Hazelnut	100/98.7	99.9/98.2	100/99.2	100/98.6	100/99.7	98.8/98.9	100/99.1	99.9/99
	Metal nut	100/98.4	100/99.6	100/98.1	100/97.0	98.7/99.5	90.8/98.1	100/97.7	99.4/98.5
	Pill	96.6/97.4	99.6/95.7	98.4/98.3	100/98.8	98.9/97.6	98.0/92.3	97.3/98.7	98.9/98.5
	Screw	98.1/99.4	97.0/99.1	98.9/99.7	100/98.0	93.9/97.6	97.8/99.4	100/99.6	97.7/99.7
	Toothbrush	100/98.7	99.5/93.0	100/99.1	98.1/99.2	100/98.1	100/92.7	100/99.4	99.7/99.2
	Transistor	100/96.3	96.7/95.4	98.5/94.3	100/97.8	93.1/90.9	88.1/94.3	98.7/95.4	100/98.2
	Zipper	99.4/98.8	98.5/98.2	98.6/98.8	100/99.2	100/98.8	67.5/93.4	99.6/98.3	98.8/98.8
	Mean	99.1/98.1	98.5/97.8	99.4/98.3	99.6/97.6	98.0/97.3	90.5/95.5	99.6/98.2	99.4/98.5

screw bags and pushpins categories, both logical and structural anomalies tend to be small in scale, and normal images often exhibit irregular arrangements. Our proposed self-distillation strategy effectively mitigates the issue of false positives caused by these subtle differences, while the shared encoder structure is optimized to capture the diverse characteristics of normal images. These design choices contribute to the superior performance of our method in detecting anomalies under these challenging conditions, highlighting its effectiveness and robustness across various demanding anomaly detection scenarios.

Beyond quantitative assessments, our qualitative findings on the MVTec LOCO AD datasets are illustrated in the last five rows of Fig. 5. As visualized in each row, the anomaly maps highlight defective regions with sharp boundaries while maintaining low false-positive rates in the surrounding areas, demonstrating strong foreground-background separation capabilities. For example, in cases such as "Splicing Connectors," "Pushpins," and "Juice Bottle," the high-response regions align almost perfectly with the ground truth annotations, achieving exceptional pixel-wise accuracy. Additionally, Table I presents notable pixel-sPRO scores on MVTec LOCO AD, where our approach consistently outperforms existing methods like PatchCore, GCAD, and THFR. Both quantitative and qualitative results demonstrate that our model achieves outstanding performance across multiple benchmarks, effectively detecting and precisely localizing anomalies at the pixel level.

TABLE III
QUANTITATIVE RESULTS ON RETINAL OCT DATASET FOR ANOMALY DETECTION.

Method	AUC	F1-score	ACC	SEN	SPE
AE [12]	77.79	85.79	78.28	94.38	41.70
VAE [21]	80.04	85.55	77.63	95.32	37.45
Ganomaly [37]	83.53	88.65	81.60	95.86	38.80
f-AnoGAN [20]	83.35	84.73	77.50	89.89	49.36
SALAD [38]	96.42	93.42	90.64	95.69	79.15
MKD [16]	96.72	94.60	91.60	97.60	73.60
RD [15]	97.64	96.40	94.60	96.40	89.20
Hetero-AE [39]	98.94	97.10	95.76	97.46	90.64
PatchCore [8]	97.52	96.96	96.61	97.65	91.25
EfficientAD [17]	98.95	97.01	96.68	98.51	91.40
Ours	99.24	98.21	97.52	98.68	92.54

Results on VisA

The performance of our anomaly detection method was rigorously evaluated using the VisA dataset, with the results summarised in Table II. Our approach achieved exceptional results in both image-level anomaly detection (measured by image-AUROC) and anomaly localization (measured by pixel-AUROC). Specifically, it achieved impressive scores of 97.5% for image-level and 99.1% for pixel-level anomaly detection,

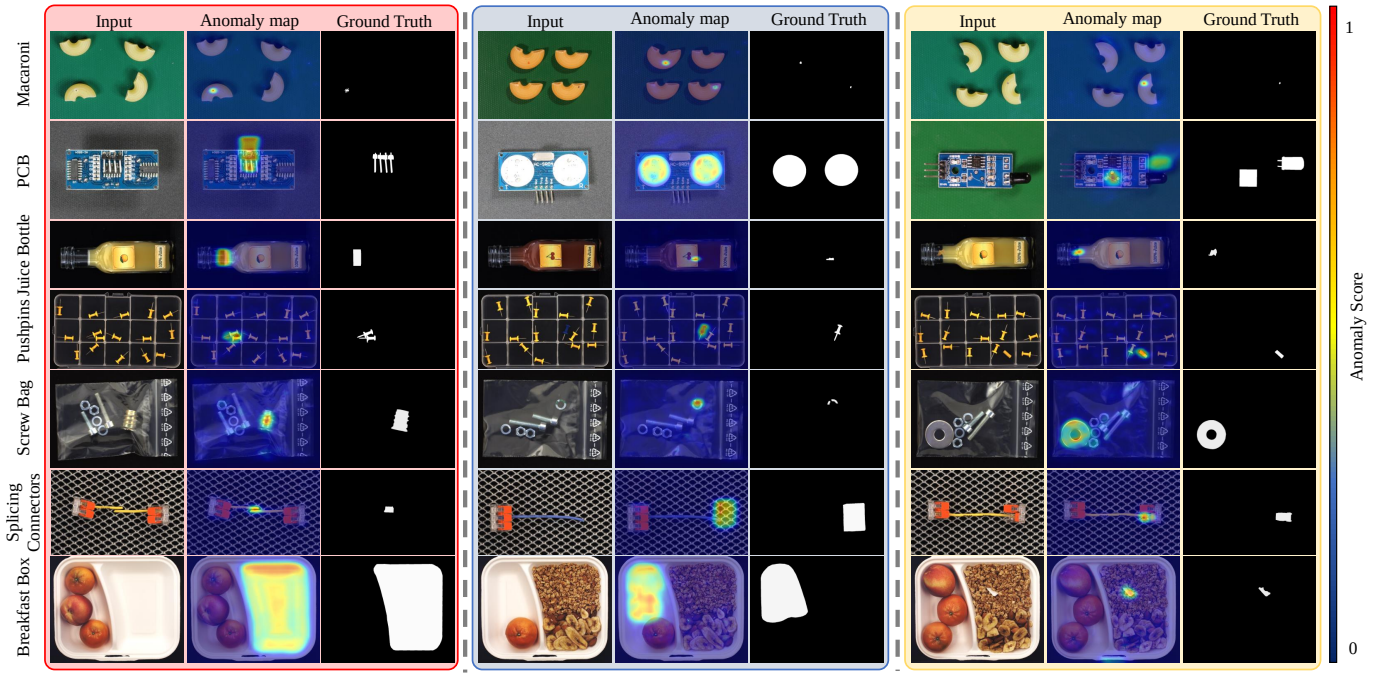


Fig. 5. Visualization of anomaly localization results on different categories of MVTec LOCO AD dataset and VisA dataset.

exceeding the current leading technology by 0.7% and 0.3%, respectively. The results of the visualization of the Visa dataset are shown in the first two rows of Fig. 5, where the accurate localization in different PCBs and Macaroni highlights the robustness and reliability of our method. These results highlight its capability to consistently detect and localize defects with high accuracy, making it an indispensable and valuable tool for effective anomaly detection in industrial applications. It is worth noting that achieving improvements in established performance metrics is challenging due to the maturity of current state-of-the-art technologies. This achievement highlights the advanced capabilities of our method in detecting anomalies accurately and efficiently.

Results on Retinal OCT Dataset

To validate the generalizability of our method, we conducted experiments on the retinal OCT dataset, which features medical images that are fundamentally different from industrial datasets. For performance evaluation, we use five key metrics commonly employed in medical applications: area under the curve (AUC), F1-score, average classification accuracy (ACC), sensitivity (SEN), and specificity (SPE). As detailed in Table III, our anomaly detection method achieves exceptional performance with an AUC of 99.24%, an F1-score of 98.21%, and an accuracy of 97.52%. These results not only surpass those of existing state-of-the-art methods but also highlight the robustness of our approach across diverse domains. Additionally, the F1-score indicates a strong balance between precision and recall, crucial for minimizing false positives and ensuring accurate anomaly detection in clinical settings. A significant factor contributing to this success is our innovative hybrid distillation strategy, which leverages self-distillation to enhance the detection of subtle anomalies. This approach effectively mitigates issues related to reconstruction quality in the logical branch, reducing false positives that can arise from minor variations. Overall,

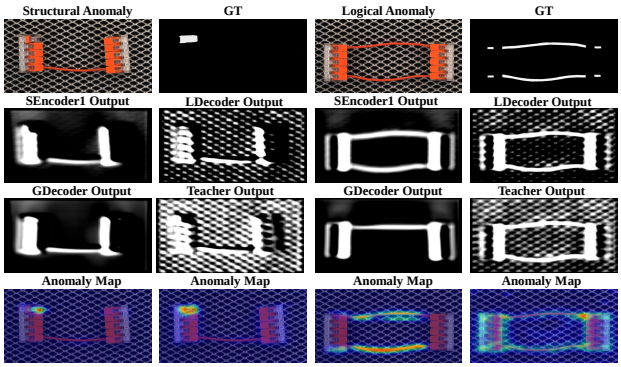


Fig. 6. Visualization results for one of the channels of our proposed approach

the consistent improvement across all metrics underscores the robustness of our method.

Results on MVTec AD

To demonstrate the superiority and practicality of our proposed method, we conducted a comparative analysis with several state-of-the-art (SOTA) methods on the well-established MVTec AD dataset. The quantitative results are presented in Table II. It's important to note that the MVTec AD dataset predominantly consists of structural anomalies, typically a single target object against a simple background. We also employed an early-stopping strategy in our experiments. The results in Table II clearly indicate that our proposed method achieves outstanding performance. Specifically, our method attains an image-level AUROC of 99.4%, demonstrating its superior capability in detecting anomalies at a high level of accuracy. Moreover, for anomaly localization, our method achieves a pixel-level AUROC of 98.5%, indicating its exceptional precision in identifying the exact locations of anomalies within images. These results underscore the effectiveness

of our method in both anomaly detection and localization tasks. The high AUROC scores achieved by our method demonstrate its exceptional ability to accurately differentiate between normal and anomalous regions. This high level of accuracy is particularly crucial for practical applications such as quality control and automated inspection systems, where reliable identification and localization of defects are essential. The integration of early-stopping, combined with our method's ability to effectively learn both object-level and fine-grained details, contributes to its superior performance in identifying and localizing anomalies. This underscores the practicality and robustness of our approach in real-world scenarios where precision and reliability are paramount.

TABLE IV
ABLATION EXPERIMENTS ON THE MVTEC LOCO AD DATASET

Baseline	Sencoder2	loss	Pixel-sPro			Image-AUROC
			Structural	Logical	Mean	
✓			0.753	0.501	0.627	0.785
✓		✓	0.812	0.673	0.742	0.830
✓	✓		0.820	0.737	0.779	0.852
✓	✓	✓	0.871	0.787	0.829	0.910

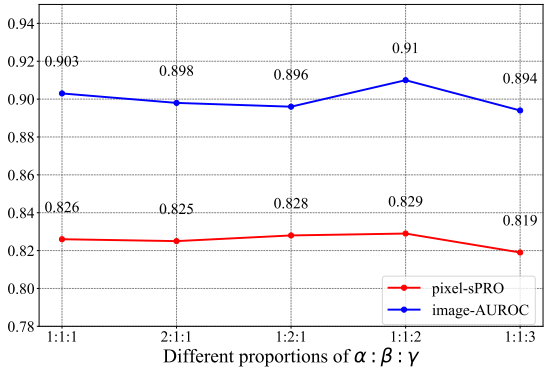


Fig. 7. Results for different proportions of the hyperparameters α , β, and γ on the MVTec LOCO AD dataset.

C. Ablation analysis

To validate the effectiveness of our method, we conducted ablation experiments to assess the contributions of self-distillation loss and the SEncoder2 module to anomaly detection performance. As shown in Table IV, incorporating self-distillation loss significantly improves all metrics, with Structural AUROC reaching 0.812, Logical AUROC 0.673, Mean AUROC 0.742, and Image-AUROC 0.830. The self-distillation loss facilitates the refinement of features by leveraging the network's own predictions as a form of "soft supervision," leading to enhanced feature representations that improve the model's ability to detect both structural and logical anomalies. When combined with the SEncoder2 module, overall performance is further enhanced, underscoring the complementary nature of these two components. The output of GDecoder, f_g (first column of the third row in Fig. 6), may lead to false alarms when directly compared with the output of the teacher network, f_t (second column of the third row in Fig. 6), due to the feature degradation during the spatial transformation. To address this issue, we develop a self-distillation loss mechanism that replaces the output of SEncoder1 with the output of the teacher network thereby improving the accuracy of anomaly detection. We further analyzed the effects of loss components by varying the hyperparameter ratios α, β, and γ on the MVTec LOCO AD dataset. This analysis focused not only on identifying the optimal combination of parameters but also on understanding the underlying reasons for performance variations across different settings. The results in Fig. 7 show that the self-distillation loss ratio is too low, the model under-utilizes the teacher guidance, leading to weaker alignment between the GDecoder's output and the teacher features. Conversely, an excessively high self-distillation ratio tends to overshadow the reconstruction-focused losses, potentially harming the model's ability to capture subtle anomalies.

In summary, the ablation experiments demonstrate that both self-distillation loss and the SEncoder2 module are essential for achieving superior anomaly detection performance.

D. Computational efficiency analysis

We compared our method with other representative approaches, focusing on computational efficiency and practical deployment. Table V provides metrics including Pixel-sPRO, parameters, FLOPs, latency (ms), and throughput (img/s), offering a comprehensive evaluation. Our method strikes an effective balance between performance and efficiency, with 35 million parameters and 234 billion FLOPs, comparable to EfficientAD's 21 million parameters and 235 billion FLOPs. Importantly, it achieves a latency of 8.2 ms and a throughput of 151 img/s, meeting real-world industrial requirements. In terms of anomaly detection, our method achieves the highest Pixel-sPRO value of 0.829, significantly outperforming EfficientAD's 0.798. In summary, our method's ability to strike a commendable balance between computational efficiency and practical applicability. Its superior performance metrics, combined with low latency and high throughput, ensure that our approach is not only robust in theory but also viable for deployment in real-time industrial scenarios.

TABLE V
COMPUTATIONAL EFFICIENCY ANALYSIS EXPERIMENTS ON THE MVTEC LOCO AD DATASET

Method	Pixel-sPRO	Number of Parameters [$\times 10^6$]	FLOPs[$\times 10^9$]	Latency[ms]	Throughput[img/s]
GCAD [28]	0.701	65	416	16	82
PatchCore [8]	0.546	150 + 8	159 + kNN	45	52
S–T [14]	0.626	26	4468	88	13
FastFlow [30]	0.568	92	85	24	66
SimpleNet [29]	0.363	73	38	18	79
EfficientAD [17]	0.798	21	235	6.5	201
Ours	0.829	35	234	8.2	151

V. CONCLUSION

Our proposed multi-task hybrid knowledge distillation method for anomaly detection and localization incorporates two essential techniques: multi-task student networks and self-distillation loss. This hybrid distillation approach minimizes false positives that often arise from image reconstruction blurring due to feature compression in the logical branch. Employing this approach, we attained state-of-the-art performance on the MVTec LOCO AD dataset and achieved competitive results on the MVTec AD and VisA datasets. Our method offers a robust and straightforward solution for addressing

the complexities of anomaly detection, demonstrating its potential applicability and effectiveness across a range of industrial scenarios. However, the performance of knowledge distillation-based methods heavily depends on the quality of the reconstructed image. Due to the presence of the bottleneck layer, the detection effectiveness may decrease if the anomalies are too small or too similar to the background. In the future, with the rise of large-scale models, general-purpose zero-shot anomaly detection driven by such models is expected to gain widespread attention and experience continuous development.

REFERENCES

[1] X. Tao, D. Zhang, W. Ma, Z. Hou, Z. Lu, and C. Adak, "Unsupervised anomaly detection for surface defects with dual-siamese network," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 11, pp. 7707–7717, 2022.

[2] Y. Zhang, X. Nie, R. He, M. Chen, and Y. Yin, "Normality learning in multispace for video anomaly detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 9, pp. 3694–3706, 2020.

[3] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *arXiv preprint arXiv:1901.03407*, 2019.

[4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[5] S. Zhang, M. Gong, Y. Xie, A. K. Qin, H. Li, Y. Gao, and Y.-S. Ong, "Influence-aware attention networks for anomaly detection in surveillance videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5427–5437, 2022.

[6] G. Wang, S. Han, E. Ding, and D. Huang, "Student-teacher feature pyramid matching for anomaly detection," *arXiv preprint arXiv:2103.04257*, 2021.

[7] J. Jiang, J. Zhu, M. Bilal, Y. Cui, N. Kumar, R. Dou, F. Su, and X. Xu, "Masked swin transformer unet for industrial anomaly detection," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 2200–2209, 2022.

[8] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 318–14 328.

[9] T. Reiss, N. Cohen, L. Bergman, and Y. Hoshen, "Panda: Adapting pretrained features for anomaly detection and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2806–2814.

[10] M. Xu, X. Zhou, X. Gao, W. He, and S. Niu, "Discriminative feature learning framework with gradient preference for anomaly detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–10, 2022.

[11] W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R. J. Radke, and O. Camps, "Towards visually explaining variational autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8642–8651.

[12] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special lecture on IE*, vol. 2, no. 1, pp. 1–18, 2015.

[13] H. Guo, L. Ren, J. Fu, Y. Wang, Z. Zhang, C. Lan, H. Wang, and X. Hou, "Template-guided hierarchical feature restoration for anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 6447–6458.

[14] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4183–4192.

[15] H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9737–9746.

[16] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution knowledge distillation for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 902–14 912.

[17] K. Batzner, L. Heckler, and R. König, "Efficientad: Accurate visual anomaly detection at millisecond-level latencies," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 128–138.

[18] J. Zhang, M. Suganuma, and T. Okatani, "Contextual affinity distillation for image anomaly detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 149–158.

[19] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "Padim: a patch distribution modeling framework for anomaly detection and localization," in *International Conference on Pattern Recognition (ICPR)*. Springer, 2021, pp. 475–489.

[20] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Medical image analysis*, vol. 54, pp. 30–44, 2019.

[21] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," *arXiv preprint arXiv:1807.02011*, 2018.

[22] X. Yan, H. Zhang, X. Xu, X. Hu, and P.-A. Heng, "Learning semantic context from normal samples for unsupervised anomaly detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 4, 2021, pp. 3110–3118.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[24] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Asymmetric student-teacher networks for industrial anomaly detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2592–2602.

[25] C. Steger, M. Ulrich, and C. Wiedemann, *Machine vision algorithms and applications[M]*. John Wiley & Sons, 2018.

[26] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14 372–14 381.

[27] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," *arXiv preprint arXiv:2005.02357*, 2020.

[28] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization," *International Journal of Computer Vision*, vol. 130, no. 4, pp. 947–969, 2022.

[29] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "Simplenet: A simple network for image anomaly detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 402–20 411.

[30] J. Yu, Y. Zheng, X. Wang, W. Li, Y. Wu, R. Zhao, and L. Wu, "Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows," *arXiv preprint arXiv:2111.07677*, 2021.

[31] V. Zavrtanik, M. Kristan, and D. Skočaj, "Draem-a discriminatively trained reconstruction embedding for surface anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 8330–8339.

[32] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 392–408.

[33] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9592–9600.

[34] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[35] T. D. Tien, A. T. Nguyen, N. H. Tran, T. D. Huy, S. Duong, C. D. T. Nguyen, and S. Q. Truong, "Revisiting reverse distillation for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 511–24 520.

[36] W. Liu, H. Chang, B. Ma, S. Shan, and X. Chen, "Diversity-measurable anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 147–12 156.

[37] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, 2019, pp. 622–637.

[38] H. Zhao, Y. Li, N. He, K. Ma, L. Fang, H. Li, and Y. Zheng, "Anomaly detection for medical images using self-supervised and translation-consistent features," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3641–3651, 2021.

[39] S. Lu, W. Zhang, H. Zhao, H. Liu, N. Wang, and H. Li, "Anomaly detection for medical images using heterogeneous auto-encoder," *IEEE Transactions on Image Processing*, 2024.