Unlocking the Potential of Reverse Distillation for Anomaly Detection

Xinyue Liu[1], **Jianyuan Wang**[2*], **Biao Leng**[1], **Shuo Zhang**[3]

[1]School of Computer Science and Engineering, Beihang University
[2]School of Intelligence Science and Technology, University of Science and Technology Beijing
[3]Beijing Key Lab of Traffic Data Analysis and Mining, School of Computer & Technology, Beijing Jiaotong University
{liuxinyue7, lengbiao}@buaa.edu.cn, wangjianyuan@ustb.edu.cn, zhangshuo@bjtu.edu.cn

Abstract

Knowledge Distillation (KD) is a promising approach for unsupervised Anomaly Detection (AD). However, the student network's over-generalization often diminishes the crucial representation differences between teacher and student in anomalous regions, leading to detection failures. To address this problem, the widely accepted Reverse Distillation (RD) paradigm designs the asymmetry teacher and student, using an encoder as teacher and a decoder as student. Yet, the design of RD does not ensure that the teacher encoder effectively distinguishes between normal and abnormal features or that the student decoder generates anomaly-free features. Additionally, the absence of skip connections results in a loss of fine details during feature reconstruction. To address these issues, we propose RD with Expert, which introduces a novel Expert-Teacher-Student network for simultaneous distillation of both the teacher encoder and student decoder. The added expert network enhances the student's ability to generate normal features and optimizes the teacher's differentiation between normal and abnormal features, reducing missed detections. Additionally, Guided Information Injection is designed to filter and transfer features from teacher to student, improving detail reconstruction and minimizing false positives. Experiments on several benchmarks prove that our method outperforms existing unsupervised AD methods under RD paradigm, fully unlocking RD's potential.

Code — https://github.com/hito2448/URD

Introduction

Anomaly Detection (AD) is one of the key tasks in industry. Due to the difficulty in obtaining anomalous images and the high cost of labeling, unsupervised Anomaly Detection has been extensively studied. Unsupervised AD uses only normal images during training, enabling the model to detect and localize anomalies in the test images. In recent years, thanks to the application of techniques such as reconstruction models, diffusion models, normalizing flow, and knowledge distillation, unsupervised AD has seen rapid advancements.

Knowledge distillation is one of the common paradigms for unsupervised AD (Bergmann et al. 2020). Similar to traditional knowledge distillation, KD-based AD methods typically rely on a teacher-student network, where an initialized student network is distilled by the pre-trained teacher network. Since the student network is trained only on normal images, it is unable to obtain the teacher network's anomaly representation ability. Therefore, during inference, the difference in feature representations between the teacher and student networks is used to determine whether there are anomalies. Early KD-based AD methods (Bergmann et al. 2020; Salehi et al. 2021; Wang et al. 2021; Zhou et al. 2022) use teacher and student networks with identical or similar architectures and data flow, leading to the student over-generalizing the teacher's anomaly representation ability. To tackle this shortfall, Reverse Distillation (RD) (Deng and Li 2022) innovatively combines the concept of knowledge distillation with feature reconstruction, using a pre-trained encoder as the teacher and a decoder as the student. The asymmetric network architectures and the reverse data flow of RD results in better anomaly localization performance.

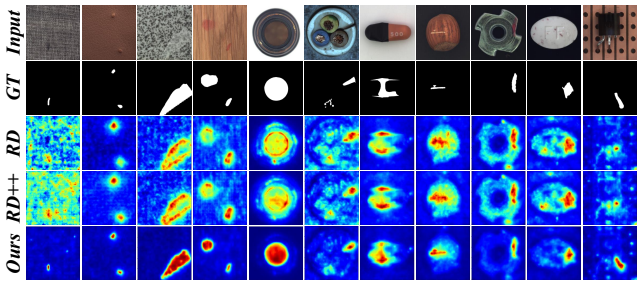

Figure 1: Anomaly localization examples. Our method reduces missed detections and false positives in RD.

Although RD is simple and effective for unsupervised AD, there exist some shortcomings in its architectural design: (1) RD's bottleneck module claims to filter out abnormal information so that the student decoder generates anomaly-free features. However, since there is only normal supervision during training, the anomaly filtering is not explicitly guaranteed. Thus, in some cases, the decoder still reconstructs features similar to the teacher encoder's, giving rise to missed detections. (2) To inject encoder features into the decoder for better detail reconstruction from high-level representations, most reconstruction networks introduce skip connection. To prevent anomaly leakage, RD, though as a feature reconstruction network, discards skip

*Corresponding author.

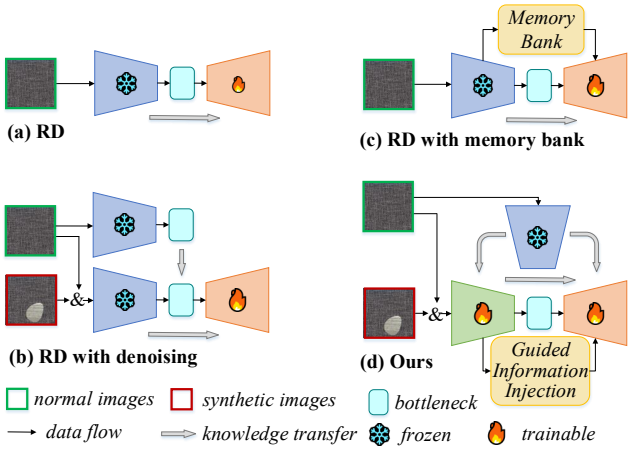

Figure 2: Schematic diagram of the framework and data flow of RD and its variants including our proposed method.

connections, limiting its ability to reconstruct fine details and leading to false positives in normal regions.

To address the above issues of RD, recent methods propose improvements in two aspects: **enriching supervision information** and **expanding reconstruction information**. For the lack of anomaly supervision, some methods integrate the denoising concept with anomaly synthesis, as shown in Figure 2 (b) (Tien et al. 2023; Jiang, Cao, and Shen 2023). For instance, the representative method RD++ (Tien et al. 2023) trains the bottleneck to reconstruct the normal feature space from the teacher encoder's abnormal features, thereby ensuring that the student decoder outputs normal features. However, this denoising strategy is based on the strong hypothesis that the features generated by the teacher encoder in anomalous regions differ from the corresponding normal features reconstructed by the student decoder. When the proportion of normal pixels in the receptive field is large, this hypothesis may not hold. For poor detail reconstruction, other methods (Guo et al. 2023; Gu et al. 2023) introduce the memory bank to store normal features of the teacher encoder, thus expand the available information of the student decoder during reconstruction, as in Figure 2 (c). However, the memory bank brings additional storage requirements, and feature search and alignment also require additional computing requirements.

To unlock the potential of the RD paradigm for unsupervised AD task, we continue to innovate in **comprehensive supervision** and **detail reconstruction**. First, to ensure the efficacy of RD to a great extent, our idea is to incorporate synthetic anomalies to explicitly denoise the student decoder's features while also distill the teacher encoder's features. By enlarging the difference between features generated in normal and anomalous regions, the teacher's anomaly sensitivity is improved. Additionally, to better reconstruct detail information in lower-level features, our intuition is to employ similarity attention to directly transfer the teacher's feature information into student, thereby straightforwardly enhancing detail reconstruction.

Unlike the previous RD framework, as illustrated in Figure 2 (d), we propose Reverse Distillation with Expert (RD-E) based on an innovative Expert-Teacher-Student Network, which leverages a frozen expert encoder to simultaneously train the teacher encoder and student decoder, enhancing their anomaly sensitivity and denoising capability. In addition, considering that skip connection may cause anomaly leakage, we design Guided Information Injection, utilizing teacher's selective information to aid the student decoder in reconstructing low-level feature details. Experimental results on widely benchmarked AD datasets demonstrate that anomaly detection and localization performance of our method surpasses that of RD and other mainstream KD-based methods, achieving SOTA.

Related Works

As one of the crucial tasks in industrial quality inspection (Bergmann et al. 2019), unsupervised Anomaly Detection (AD) has earned increasing attention in recent years. Early methods in unsupervised AD often relies on generative models (Bergmann et al. 2018; Akcay, Atapour-Abarghouei, and Breckon 2019; Tang et al. 2020; Liu et al. 2023a; Zhang et al. 2023a), where the models are trained on normal samples to learn how to reconstruct them and the reconstruction error is used for inference. Other methods employ parametric density estimation (Defard et al. 2021; Gudovskiy, Ishizaka, and Kozuka 2022; Hyun et al. 2024; Zhou et al. 2024), where the parameters of normal distribution are calculated, and anomalies are detected based on how well the samples fit this distribution. Additionally, many methods incorporated pre-trained models (Liu et al. 2023b; Li et al. 2023) and memory banks (Roth et al. 2022; Bae, Lee, and Kim 2023), comparing the input images to stored normal features. Recently, synthetic anomalies have become a hot topic (Li et al. 2021; Lin and Yan 2024), with external datasets (Zavrtanik, Kristan, and Skočaj 2021) or diffusion models (Zhang, Xu, and Zhou 2024) being used to generate anomalies similar to real-world scenarios, thus aiding unsupervised AD.

Besides, knowledge distillation (KD) based on teacher-student networks has recently been applied to unsupervised AD (Bergmann et al. 2020; Li et al. 2024), using differences in representation between the two networks to identify anomalies. A major concern in KD-based AD is student's over-generalization to teacher's anomaly representations. Some methods (Salehi et al. 2021; Wang et al. 2021; Rudolph et al. 2023; Liu et al. 2024) address this by using asymmetry teacher and student networks to differentiate their representation abilities. Reverse Distillation (RD) (Deng and Li 2022) follows this idea, proposing reverse network architecture and data flow. Recently, several improvements to RD have been explored (Tien et al. 2023; Guo et al. 2023; Gu et al. 2023; Zhang, Suganuma, and Okatani 2024).

Revisiting Reverse Distillation

RD (Deng and Li 2022) is a widely adopted unsupervised AD paradigm based on KD. The primary components of RD include a pre-trained teacher encoder E, a one-class bottleneck embedding (OCBE) module, and a student decoder D.

During training, only normal images are used as input.

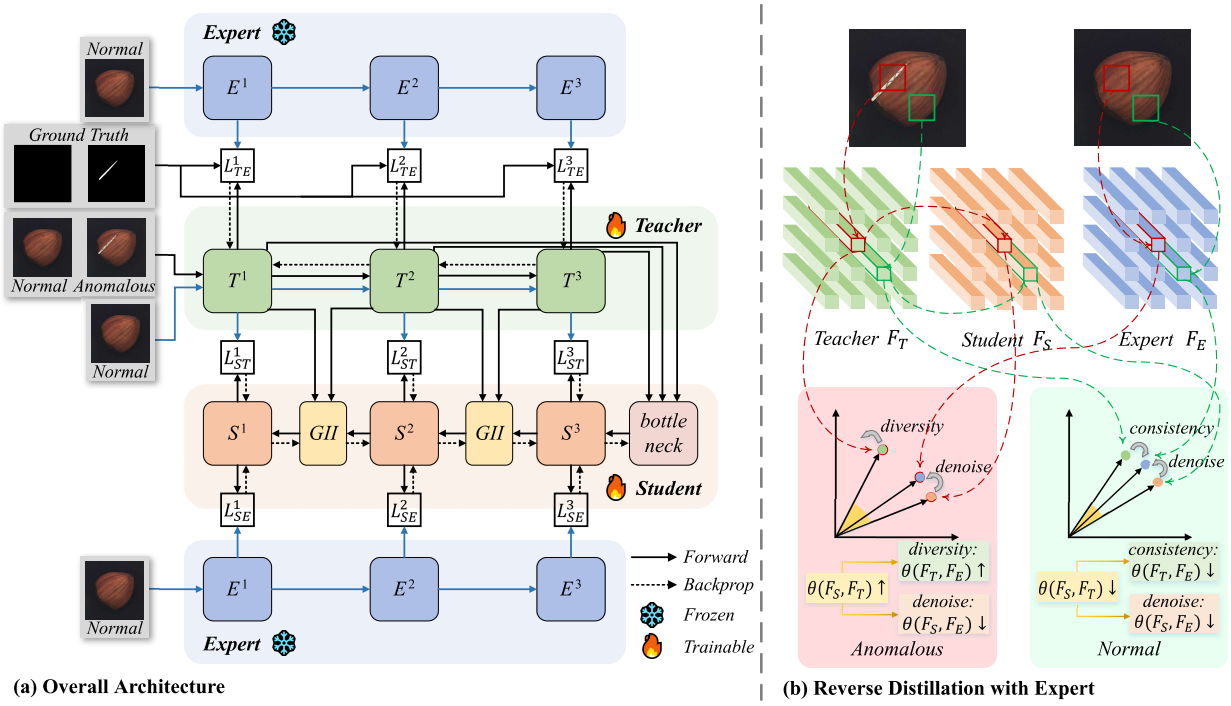

Figure 3: Overview of our proposed method. (a) shows the overall architecture and training process of our designed Expert-Teacher-Student Network, where the expert is frozen and the teacher and student are trainable. Our proposed Guided Information Injection module is inserted between the two blocks of the student. (b) shows how to distill the teacher and student with the expert. The impact of distillation on the teacher and student features is visually represented. Through the two sub-tasks: making the teacher encoder more sensitive to anomalies and better denoising the student features, differences between teacher and student features are achieved in anomalous regions, while similarities in normal regions are maintained.

The teacher network is frozen, while the bottleneck and student networks are trainable. The OCBE module compresses multi-scale patterns into a low-dimensional space. The compact embedding is then fed into the student network to reconstruct the teacher network's features.

During inference, the teacher network can capture abnormal features that deviate from the distribution of normal samples. The OCBE module, by generating compact features, prevents these abnormal perturbations from being input into the student network. Therefore, the student network can generate anomaly-free features regardless of whether normal or abnormal samples are input. The discrepancy in feature reconstruction, measured by their similarity, is used to detect and localize anomalies.

However, the design of RD still has some limitations that affect its anomaly detection performance:

(1) Miss Detection Issue: The effectiveness of RD relies on two key premises, each with specific requirements for the teacher and student. Firstly, the teacher should capture anomalies by generating abnormal features that differ from normal ones in anomalous regions. RD assumes that this premise is always valid. However, in some cases, such as when the anomaly is small and normal pixels dominate the receptive field, this assumption may not hold true. Secondly, the student is promised to generate anomaly-free features. Since OCBE essentially performs only a downsampling operation, the generated features used as student input are not guaranteed to be compact and may still contain abnormal information. Additionally, the multi-layer convolutional student decoder has strong generalization capabilities, which means that even if only trained on normal samples, it may still generate abnormal features similar to those of the teacher due to over-generalization. Consequently, the inability to meet these premises results in insufficient difference between the features of teacher and student in anomalous regions. Therefore, some anomalies are not detected.

(2) False Positive Issue: Since the teacher encoder performs multi-step downsampling and multi-layer convolution, the output high-level features lose lots of details compared with low-level features. Directly using high-level features for low-level feature reconstruction results in reconstruction errors. Most of previous reconstruction networks utilize skip connections to directly pass encoder features to the corresponding decoder layers. However, for RD, this operation may introduce abnormal information from the teacher encoder into the student decoder, making it difficult to generate anomaly-free features. To overcome this challenge, RD designes MFF, which fuses multiple layers of encoder features as the input of the decoder. Although MFF allows low-level features to be included in generating the decoder input, it still downsamples these features to a smaller scale before feature fusion. Hence, some useful detail infor-

mation is lost, which causes notable discrepancies between student's reconstructed features and teacher's features even in normal regions, thereby raising the false positive rate.

Method

The overall architecture of our proposed method is illustrated in Figure 3 (a). Based on the original teacher-student framework of RD, we design an Expert-Teacher-Student (E-T-S) Network, which retains the design of the teacher, bottleneck, and student from RD. The teacher encoder T is a WideResNet50 (Zagoruyko and Komodakis 2016) pre-trained on ImageNet (Deng et al. 2009). The bottleneck, named OCBE by RD, includes Multi-scale Feature Fusion (MFF) and One-Class Embedding (OCE) modules. The student decoder S is a symmetric network with T, differing in that it replaces downsampling by upsampling. Additionally, S includes Guided Information Injection (GII) to incorporate information from encoder. Besides, we innovatively introduce an expert network E with the same architecture and initial parameters as T.

During training, different from RD, the teacher, bottleneck, and student in E-T-S Network are all trainable. The teacher uses a separate optimizer, while the bottleneck and student share the same optimizer (with the bottleneck being considered part of the student in the following sections). During inference, the frozen teacher and student are used for anomaly detection and localization.

Reverse Distillation with Expert

The teacher network's ability to perceive anomalies and the student network's capacity to generate anomaly-free features are prerequisites for RD. Previous RD and its variants do not meet both of these two conditions, and trigger missed detection issue. To tackle this problem, we propose to introduce an expert network to distill both the teacher and the student at the same time, and ensure the distillation process covers **enhancing the teacher's anomaly sensitivity** and **denoising the student's features**, as in Figure 3 (b). The teacher network is optimized to be more sensitive to anomalies and capable of generating differentiated abnormal and normal features. Simultaneously, the student network is trained with a denoising strategy to ensure normal features are generated even when anomalous samples are input. This dual strategy, based on the introduction of expert network, ensures that the features of teacher and student are similar in normal regions and dissimilar in anomalous regions, which enables effective anomaly detection and localization.

For each normal image I_n in the training set, anomaly synthesis operation is performed to generate a corresponding synthetic anomalous image I_a. Here, we follow DRÆM (Zavrtanik, Kristan, and Skočaj 2021) for synthesizing anomalies with a Perlin noise generator and the Describable Textures Dataset (Cimpoi et al. 2014). The teacher T receives a pair of images $I = \{I_n, I_a\}$ as input and outputs three layers of features: $F_T^n = \{F_T^{n1}, F_T^{n2}, F_T^{n3}\} = T(I_n)$ and $F_T^a = \{F_T^{a1}, F_T^{a2}, F_T^{a3}\} = T(I_a)$. The student S takes the features from the teacher network as input and reconstructs the corresponding three features: $F_S^n = \{F_S^{n1}, F_S^{n2}, F_S^{n3}\} = S(F_T^n)$ and $F_S^a = \{F_S^{a1}, F_S^{a2}, F_S^{a3}\} = S(F_T^a)$. The expert E, which takes normal images as input, produces the features that correspond to those: $F_E^n = \{F_E^{n1}, F_E^{n2}, F_E^{n3}\} = E(I_n)$.

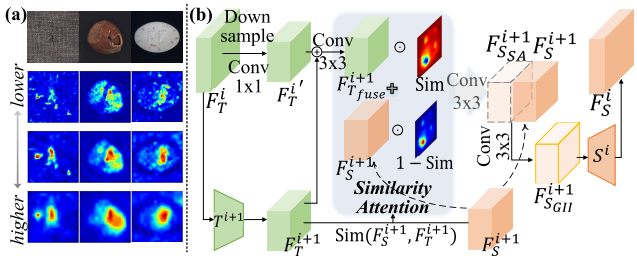

Figure 4: (a) Cosine distance maps between features of teacher and student. (b) Guided Information Injection.

To enhance the teacher's sensitivity to anomalies, we explicitly guide the teacher's feature extraction process using ground truth anomaly masks M_{gt}. We maintain high cosine similarity for normal regions. In the meantime, by increasing the cosine distances between the teacher's abnormal features and the expert's normal features in anomalous regions, the teacher network is optimized to generate differentiated abnormal and normal features. The teacher's training loss $\mathcal{L}_{TE}$ is calculated using L1 distance as

$$D_{TE}^{n/a^i}(h,w) = 1 - \frac{F_T^{n/a^i}(h,w) \cdot F_E^{ni}(h,w)}{\|F_T^{n/a^i}(h,w)\| \|F_E^{ni}(h,w)\|} \tag{1}$$

$$\mathcal{L}_{TE}^{n/a} = \sum_{i=1}^{3} \{ \frac{1}{H_i W_i} \sum_{h=1}^{H_i} \sum_{w=1}^{W_i} |D_{TE}^{n/a^i}(h,w) - M_{gt}^i| \} \tag{2}$$

$$\mathcal{L}_{TE} = \mathcal{L}_{TE}^n + \mathcal{L}_{TE}^a \tag{3}$$

where H_i and W_i represent the height and width of the output feature of the i-th encoding block. M_{gt}^i is obtained by downsampling M_{gt} to align the size of F_T^i.

To denoise the student's features, we use both the teacher and expert networks to guide the student network, ensuring that the student network generates normal features. To be specific, the student network aims at reconstructing the normal features of the teacher and expert networks whether the input images are normal or anomalous, which is optimized based on cosine similarity with $\mathcal{L}_S$ calculated as

$$f = \mathcal{F}(F) \tag{4}$$

$$\mathcal{L}_{SE/ST}^i = (1 - \frac{f_S^{ni} \cdot f_{E/T}^{n\,i}}{\|f_S^{ni}\| \|f_{E/T}^{n\,i}\|}) + (1 - \frac{f_S^{ai} \cdot f_{E/T}^{n\,i}}{\|f_S^{ai}\| \|f_{E/T}^{n\,i}\|}) \tag{5}$$

$$\mathcal{L}_S = \sum_{i=1}^{3} (L_{SE}^i + L_{ST}^i) \tag{6}$$

where $\mathcal{F}$ is the flatten operation introduced in ReContrast (Guo et al. 2024).

Guided Information Injection

Considering that: (1) Higher-level features contain less texture details, making detail reconstruction less critical. (2)

	Forward Distillation			Reverse Distillation				
	STPM	DeSTSeg	HypAD	RD	RD++	THFR	MemKD	Ours
Texture Average	-	99.1/-	-	99.7/99.9	**99.8**/99.9	99.7/-	**99.8**/-	**99.8/100**
Object Average	-	98.3/-	-	98.3/99.3	98.6/99.4	98.9/-	**99.5**/-	98.9/**99.6**
Total Average	95.5/-	98.6/-	99.2/99.5	98.8/99.5	99.0/99.6	99.2/-	**99.6**/-	99.2/**99.7**

Table 1: Image-level anomaly detection results I-AUC/I-AP (%) on MVTec AD with the best in bold.

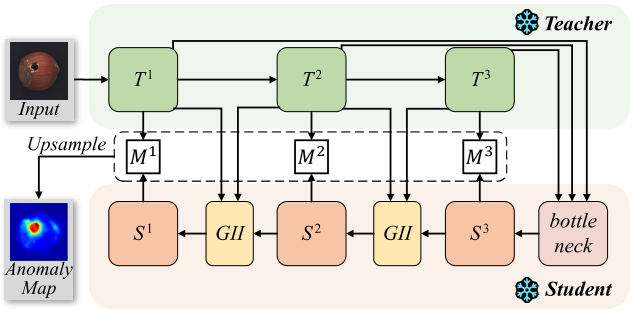

Figure 5: Inference procedure of our proposed method. The expert is removed and both teacher and student are frozen.

The shorter generation path for higher-level features naturally leads to better reconstruction quality. The distance maps calculated by the cosine similarity of higher-level features in Figure 4 (a) are therefore believed to effectively locate anomalies and highlight anomalous regions.

Inspired by this, we propose Guided Information Injection (GII), as shown in Figure 4 (b). By leveraging similarity-based attention from higher layers to guide the information injection from encoder to decoder, GII not only directly addresses the issue of the lack of low-level information during reconstruction, but also filters out most of the anomalous information, preventing anomaly leakage. As a result, detail information is introduced into the decoder in a more controlled and softer manner compared to traditional skip connections.

Specifically, GII is inserted before the student decoder blocks S^1 and S^2. For the GII module before S^i, the input consists of the output features F_T^i, F_T^{i+1}, and F_S^{i+1} from T^i, T^{i+1}, and S^{i+1}, respectively. First, F_T^i and F_T^{i+1} are adjusted in dimension and combined to obtain the multi-scale fused feature $F_{T_{fuse}}^{i+1}$. Then, the cosine similarity between the higher-level features $\mathrm{Sim}(F_T^{i+1}, F_S^{i+1})$ (hereafter referred to as Sim) is calculated, where smaller values indicate a higher likelihood of anomalies. Finally, Sim is used to control the proportion of the fused feature $F_{T\ fuse}^{i+1}$ from teacher encoder, and a feature with enriched details $F_{S_{SA}}^{i+1}$ is obtained. The original feature F_S^{i+1} and the detail-enriched feature $F_{S_{SA}}^{i+1}$ are concatenated and passed to the decoder block S^i for subsequent reconstruction, outputting the final feature F_S^i with injected detail information. For more detailed calculations, see Figure 4 (b).

Inference

Figure 5 illustrates the inference process. During inference, the expert network E is removed, ensuring that our method does not increase storage and computational overhead. The approach for anomaly scoring follows RD. For anomaly localization, the score map is obtained by summing the cosine distance maps between the three-layer features of the teacher T and the student S which are upsampled to the input image size. For anomaly detection, the image-level anomaly score is represented by the maximum value in the score map.

Experiments

Experimental Setup

Datasets We conduct our experiments primarily on **MVTec AD** (Bergmann et al. 2019) containing 5354 images across 15 categories, **MPDD** (Jezek et al. 2021) containing 1346 images across 6 categories, and **BTAD** (Mishra et al. 2021), which includes 2540 images across 3 categories. All datasets have only normal images in the training set, while have both normal and anomalous images in the test set.

Implementation Details A separate detection model is trained for each category. During both training and inference, all images are resized to 256×256. The training batch size is 16, with an early stopping strategy for a maximum of 10k iterations. Consistent with RD, the student decoder is optimized using an Adam optimizer with a learning rate of 0.005, while the teacher encoder is trained with another one at a learning rate of 0.0001. During inference, the anomaly maps are smoothed using a Gaussian filter with $\sigma = 4$.

Evaluation Metrics For anomaly detection, the used evaluation metrics are area under the receiver operating characteristic (AUROC) and average precision (AP). For anomaly localization, in addition to AUROC and AP, we also report per-region-overlap (PRO) (Bergmann et al. 2020).

Main Results

To demonstrate the superiority, we compare our method with various KD-based unsupervised AD methods, including STPM (Wang et al. 2021), DeSTSeg (Zhang et al. 2023b), and HypAD (Li et al. 2024) under Forward Distillation (FD) paradigm, as well as RD (Deng and Li 2022), RD++ (Tien et al. 2023), THFR (Guo et al. 2023), and MemKD (Gu et al. 2023) under Reverse Distillation paradigm. For a fairer comparison, we retrain the main comparison methods RD and RD++ under the same environment with our method.

	Category	Forward Distillation			Reverse Distillation				
		STPM	DeSTSeg	HypAD	RD	RD++	THFR	MemKD	Ours
Textures	Carpet	98.8/-/95.8	96.1/72.8/93.6	-/-/92.7	99.3/67.2/97.9	99.2/63.9/97.7	99.2/-/97.7	99.1/-/97.5	**99.6/83.0/98.3**
	Grid	99.0/-/96.6	99.1/61.5/96.4	-/-/99.7	99.3/**50.2/97.7**	99.3/49.5/**97.7**	99.3/-/**97.7**	99.2/-/96.9	**99.4**/50.1/97.5
	Leather	99.3/-/98.0	99.7/75.6/99.0	-/-/99.9	99.5/52.6/99.2	99.4/51.4/99.2	99.4/-/99.2	99.5/-/99.2	**99.7/70.0/99.3**
	Tile	97.4/-/92.1	98.0/90.0/95.5	-/-/99.8	95.8/53.8/91.1	96.4/56.2/92.1	95.5/-/90.8	95.7/-/91.1	**99.2/94.8/96.8**
	Wood	97.2/-/93.6	97.7/81.9/96.1	-/-/95.3	95.3/51.5/93.2	95.7/51.8/93.2	95.3/-/93.3	95.3/-/91.2	**98.1/80.2/95.2**
	Average	98.3/-/95.2	98.1/76.4/96.1	-/-/97.5	97.8/55.1/95.8	98.0/54.6/96.0	97.7/-/95.7	97.8/-/95.2	**99.2/75.6/97.4**
Objects	Bottle	98.8/-/95.1	99.2/90.3/96.6	-/-/100	98.8/78.4/96.9	98.7/80.0/96.9	98.9/-/97.2	98.8/-/97.1	**99.3/91.6/97.9**
	Cable	95.5/-/87.7	97.3/60.4/86.4	-/-/93.3	97.8/59.6/92.6	98.4/63.6/93.9	98.5/-/94.8	98.3/-/93.4	**98.7/73.1/94.9**
	Capsule	98.3/-/92.2	99.1/56.3/94.2	-/-/96.9	98.8/46.6/96.4	**98.9**/47.4/96.5	98.7/-/95.9	98.8/-/96.2	**98.9/50.5/96.8**
	Hazelnut	98.5/-/94.3	99.6/88.4/97.6	-/-/99.7	**99.2**/67.9/96.0	**99.2**/66.5/**96.3**	99.2/-/96.2	99.1/-/95.7	**99.2/68.0**/96.1
	Metal_nut	97.6/-/94.5	98.6/93.5/95.0	-/-/98.0	97.5/81.8/93.3	98.0/**83.9**/93.2	97.4/-/90.5	97.2/-/90.8	**98.4**/83.7/**93.7**
	Pill	97.8/-/96.5	98.7/83.1/95.3	-/-/98.4	98.4/80.2/96.9	98.4/79.6/97.1	98.0/-/96.4	98.3/-/96.6	**98.7/83.7/97.5**
	Screw	98.3/-/93.0	98.5/58.7/92.5	-/-/95.6	**99.6**/54.9/**98.4**	**99.6/55.5**/98.3	99.5/-/98.2	**99.6**/-/98.2	**99.6**/48.8/98.3
	Toothbrush	98.9/-/92.2	99.3/75.2/94.0	-/-/99.9	99.1/53.1/94.6	99.1/56.3/94.5	99.2/-/94.7	98.9/-/92.2	**99.3/68.5/95.5**
	Transistor	82.5/-/69.5	89.1/64.8/85.7	-/-/100	93.1/55.9/79.6	94.4/58.3/82.8	95.9/-/85.9	96.4/-/85.3	**97.5/70.3/90.2**
	Zipper	98.5/-/95.2	99.1/85.2/97.4	-/-/94.7	**98.9**/61.5/**96.8**	**98.9**/60.5/96.4	98.7/-/96.6	98.5/-/95.9	**98.9/69.3/96.8**
	Average	96.5/-/90.9	97.9/75.6/93.5	-/-/97.6	98.1/64.0/94.2	98.4/65.2/94.6	98.4/-/94.6	98.4/-/94.1	**98.9/70.8/95.8**
	Total Average	97.0/-/92.1	97.9/75.8/94.4	98.0/62.5/97.6	98.0/61.0/94.7	98.2/61.6/95.1	98.2/-/95.0	98.2/-/94.5	**99.0/72.4/96.3**

Table 2: Pixel-level anomaly localization results P-AUC/P-AP/P-PRO (%) on MVTec AD with the best KD-based results underlined and the best RD-based results in bold.

Category	Bracket Black	Bracket Brown	Bracket White	Connector	Metal Plate	Tubes	Average
RD	98.1/6.2/92.1	97.2/25.7/95.4	99.4/15.6/97.8	99.5/64.2/96.9	99.1/93.9/96.2	99.2/76.0/97.6	98.7/46.9/96.0
RD++	98.2/9.8/92.8	97.1/25.6/94.9	**99.5**/12.8/97.2	99.3/61.3/96.0	99.1/93.3/96.1	99.2/74.8/97.4	98.7/46.3/95.7
MemKD	97.8/10.7/94.5	96.3/20.5/95.2	98.8/15.9/97.3	99.4/60.6/96.4	99.1/94.2/95.2	99.2/74.0/97.3	98.4/46.1/95.9
Ours	**98.7/21.4/96.0**	**98.6/30.3/96.7**	99.4/**17.1/98.2**	**99.6/73.3/97.7**	**99.2/95.3/96.7**	**99.4/77.4/98.1**	**99.2/52.5/97.2**

Table 3: Pixel-level anomaly localization results P-AUC/P-AP/P-PRO (%) on MPDD with the best in bold.

Category	RD	RD++	Ours
Class 01	96.7/50.0/77.7	96.1/48.3/71.7	**97.2/55.0/78.6**
Class 02	96.8/65.9/66.5	96.5/60.1/**69.4**	**97.4/78.2**/66.9
Class 03	99.1/53.5/87.3	99.7/59.2/87.2	**99.8/62.5/90.0**
Average	97.5/56.5/77.2	97.4/55.9/76.1	**98.1/65.2/78.5**

Table 4: Pixel-level anomaly localization results P-AUC/P-AP/P-PRO (%) on BTAD with the best in bold.

Anomaly Detection Table 1 shows the image-level anomaly detection results on MVTec AD (detailed per-category results are provided in the supplementary materials). For AUC, our method is comparable to the leading KD-based AD methods in overall average. Regarding AP, our method achieves SOTA performance, with an average of 99.7% over all categories.

Anomaly Localization We conduct the quantitative comparison of anomaly localization results on MVTec AD in Table 2. Our method surpasses previous KD-based SOTA in pixel-level AUC, achieving 99.0%. While our method ranks second in pixel-level AP and PRO metrics with 72.4% and 96.3%, it represents the best performance within the RD paradigm. Qualitative visual results are shown in Figure 1.

Furthermore, we extend the quantitative comparison to MPDD and BTAD datasets. Tables 3 and 4 respectively present the anomaly localization results over all categories on MPDD and BTAD. Our method achieves the best performance in all metrics on the both datasets compared with other RD-based methods, further validating its localization capability.

Ablation Analysis

Ablation Study on Network Composition Our proposed method primarily includes two innovative components: the design of Reverse Distillation with Expert for distillation supervision innovation and the design of Guided Information Injection for network detail optimization. To demonstrate the effectiveness and necessity of the components, we conduct ablation experiments on MVTec AD, MPDD, and BTAD datasets, as shown in Table 5. The quantitative results indicate that when both RD-E and GII are applied simultaneously, the method achieves the best localization results.

In addition, Figure 6 illustrates the qualitative comparison results, where the baseline refers to the standard RD.

Expert	GII	MVTec AD			MPDD			BTAD		
		P-AUC	P-AP	P-PRO	P-AUC	P-AP	P-PRO	P-AUC	P-AP	P-PRO
-	-	98.02	61.01	94.70	98.75	46.95	95.99	97.53	55.45	77.17
✓	-	98.69	71.80	96.11	98.76	45.75	96.06	97.97	63.63	77.78
-	✓	98.17	60.20	94.77	99.14	48.95	97.17	97.83	59.14	78.23
✓	✓	98.97	72.37	96.31	99.14	52.46	97.25	98.11	65.24	78.48

Table 5: Ablation localization results (%) of network composition on MVTec AD, MPDD, and BTAD.

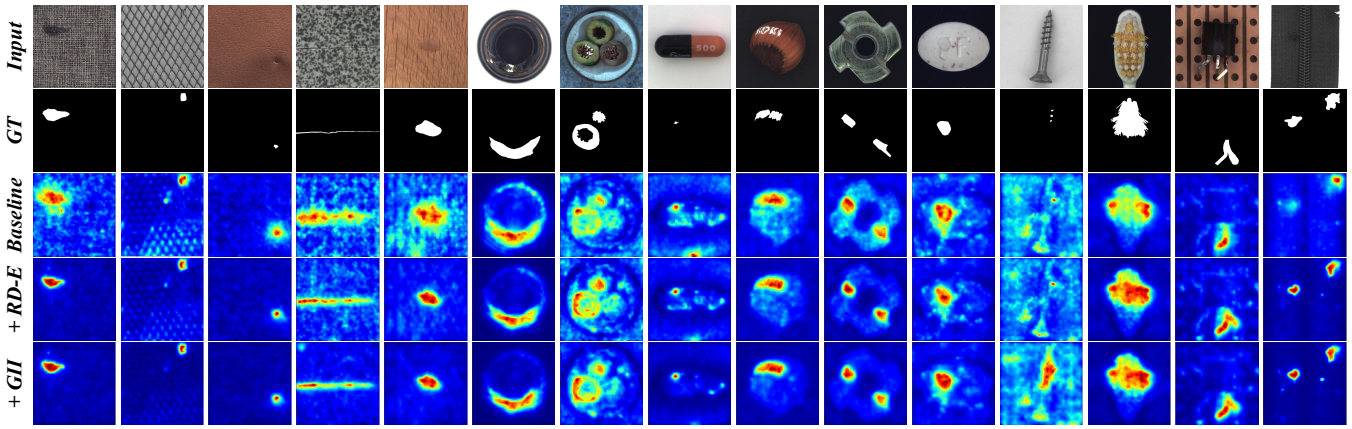

Figure 6: Visualization of ablation study on network composition. From top to bottom: the input image, the ground truth masks, the output anomaly maps of Baseline (RD), Baseline+RD-E, and Baseline+RD-E+GII (Ours).

Den	Sen	I-AUC	P-AUC	I-AP	P-AP	P-PRO
Teacher-Student Network						
-	-	98.77	98.02	99.52	61.01	94.70
✓	-	98.62	98.29	99.47	61.56	95.39
Expert-Teacher-Student Network						
✓	✓	98.72	98.69	99.48	71.80	96.11

Table 6: Ablation study results (%) of RD-E on MVTec AD. (Den: Denoising. Sen: Sensitivity.)

It is evident that incorporating RD-E significantly enhances anomaly localization capabilities, reducing the missed detection rate. Furthermore, with the introduction of GII, background noise in obtained anomaly maps is greatly reduced, leading to a lower false positive rate. These findings align well with our previous analysis.

Ablation Study on Reverse Distillation with Expert In Table 6, we compare the results on MVTec AD between using only the teacher encoder for the student decoder's feature **den**oising (Den) and introducing an expert network that enhances the teacher's anomaly **sen**sitivity while also **den**oising the student's features (Sen+Den). The results show a significant improvement in anomaly localization when the expert network is added to aid the distillation.

Ablation Study on Guided Information Injection Table 7 presents the results of ablation experiments related to the GII module on MVTec AD. The "+SC" row indicates the absence of similarity attention, where $F^{i+1}_{S_{SA}} = F^{i+1}_{T_{fuse}}$. The "+SA" row shows the results when similarity attention is introduced to filter features. The overall results highlight the effectiveness of GII and underscores the importance of the similarity attention mechanism within it.

		I-AUC	P-AUC	I-AP	P-AP	P-PRO
w/o GII		98.72	98.69	99.48	71.80	96.11
w/ GII	+ SC	98.96	98.86	99.56	71.14	96.19
	+ SA	99.22	98.97	99.74	72.37	96.31

Table 7: Ablation study results (%) of GII on MVTec AD. (SC: Naive skip connection. SA: Similarity attention.)

Conclusion

In this paper, we first improve Reverse Distillation with Expert for unsupervised AD. Building on the RD paradigm, we introduce an expert network that distills both the teacher and student networks, ensuring the effectiveness of RD by enhancing the teacher's sensitivity to anomalies and maintaining the student's ability to produce normal features. Besides, to address the challenge of detail reconstruction, we design Guided Information Injection, which uses high-level feature similarity as attention to guide the injection of teacher's features into the student. With these innovations, our method effectively reduces missed detections and false positives in RD, as confirmed by experimental results.

Acknowledgements

This work is supported by National Natural Science Foundation of China 62402035.

References

Akcay, S.; Atapour-Abarghouei, A.; and Breckon, T. P. 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*, 622–637. Springer.

Bae, J.; Lee, J.-H.; and Kim, S. 2023. Pni: industrial anomaly detection using position and neighborhood information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6373–6383.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 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*, 9592–9600.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2020. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4183–4192.

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

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Defard, T.; Setkov, A.; Loesch, A.; and Audigier, R. 2021. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, 475–489. Springer.

Deng, H.; and 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*, 9737–9746.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Gu, Z.; Liu, L.; Chen, X.; Yi, R.; Zhang, J.; Wang, Y.; Wang, C.; Shu, A.; Jiang, G.; and Ma, L. 2023. Remembering Normality: Memory-guided Knowledge Distillation for Unsupervised Anomaly Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16401–16409.

Gudovskiy, D.; Ishizaka, S.; and Kozuka, K. 2022. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 98–107.

Guo, H.; Ren, L.; Fu, J.; Wang, Y.; Zhang, Z.; Lan, C.; Wang, H.; and Hou, X. 2023. Template-guided Hierarchical Feature Restoration for Anomaly Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6447–6458.

Guo, J.; Jia, L.; Zhang, W.; Li, H.; et al. 2024. Recontrast: Domain-specific anomaly detection via contrastive reconstruction. *Advances in Neural Information Processing Systems*, 36.

Hyun, J.; Kim, S.; Jeon, G.; Kim, S. H.; Bae, K.; and Kang, B. J. 2024. ReConPatch: Contrastive patch representation learning for industrial anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2052–2061.

Jezek, S.; Jonak, M.; Burget, R.; Dvorak, P.; and Skotak, M. 2021. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*, 66–71. IEEE.

Jiang, Y.; Cao, Y.; and Shen, W. 2023. A masked reverse knowledge distillation method incorporating global and local information for image anomaly detection. *Knowledge-Based Systems*, 280: 110982.

Li, C.-L.; Sohn, K.; Yoon, J.; and Pfister, T. 2021. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9664–9674.

Li, H.; Chen, Z.; Xu, Y.; and Hu, J. 2024. Hyperbolic Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17511–17520.

Li, H.; Hu, J.; Li, B.; Chen, H.; Zheng, Y.; and Shen, C. 2023. Target before shooting: Accurate anomaly detection and localization under one millisecond via cascade patch retrieval. *arXiv preprint arXiv:2308.06748*.

Lin, J.; and Yan, Y. 2024. A Comprehensive Augmentation Framework for Anomaly Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8742–8749.

Liu, T.; Li, B.; Du, X.; Jiang, B.; Geng, L.; Wang, F.; and Zhao, Z. 2023a. Fair: frequency-aware image restoration for industrial visual anomaly detection. *arXiv preprint arXiv:2309.07068*.

Liu, X.; Wang, J.; Leng, B.; and Zhang, S. 2024. Dual-modeling decouple distillation for unsupervised anomaly detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5035–5044.

Liu, Z.; Zhou, Y.; Xu, Y.; and Wang, Z. 2023b. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20402–20411.

Mishra, P.; Verk, R.; Fornasier, D.; Piciarelli, C.; and Foresti, G. L. 2021. VT-ADL: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, 01–06. IEEE.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2022. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328.

Rudolph, M.; Wehrbein, T.; Rosenhahn, B.; and Wandt, B. 2023. Asymmetric student-teacher networks for industrial anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2592–2602.

Salehi, M.; Sadjadi, N.; Baselizadeh, S.; Rohban, M. H.; and Rabiee, H. R. 2021. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14902–14912.

Tang, T.-W.; Kuo, W.-H.; Lan, J.-H.; Ding, C.-F.; Hsu, H.; and Young, H.-T. 2020. Anomaly detection neural network with dual auto-encoders GAN and its industrial inspection applications. *Sensors*, 20(12): 3336.

Tien, T. D.; Nguyen, A. T.; Tran, N. H.; Huy, T. D.; Duong, S.; Nguyen, C. D. T.; and Truong, S. Q. 2023. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24511–24520.

Wang, G.; Han, S.; Ding, E.; and Huang, D. 2021. Student-Teacher Feature Pyramid Matching for Anomaly Detection. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, 306. BMVA Press.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zavrtanik, V.; Kristan, M.; and Skočaj, D. 2021. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8330–8339.

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

Zhang, X.; Li, N.; Li, J.; Dai, T.; Jiang, Y.; and Xia, S.-T. 2023a. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6782–6791.

Zhang, X.; Li, S.; Li, X.; Huang, P.; Shan, J.; and Chen, T. 2023b. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3914–3923.

Zhang, X.; Xu, M.; and Zhou, X. 2024. RealNet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16699–16708.

Zhou, Q.; He, S.; Liu, H.; Chen, T.; and Chen, J. 2022. Pull & push: Leveraging differential knowledge distillation for efficient unsupervised anomaly detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhou, Y.; Xu, X.; Song, J.; Shen, F.; and Shen, H. T. 2024. MSFlow: Multiscale Flow-Based Framework for Unsupervised Anomaly Detection. *IEEE Transactions on Neural Networks and Learning Systems*.