

---

# Fence off Anomaly Interference: Cross-Domain Distillation for Fully Unsupervised Anomaly Detection

---

**Xinyue Liu**

School of Computer Science and Engineering  
Beihang University  
liuxinyue7@buaa.edu.cn

**Jianyuan Wang**

School of Intelligence Science and Technology  
University of Science and Technology Beijing

**Biao Leng**

School of Computer Science and Engineering  
Beihang University

**Shuo Zhang**

School of Computer & Technology  
Beijing Jiaotong University

## Abstract

Fully Unsupervised Anomaly Detection (FUAD) is a practical extension of Unsupervised Anomaly Detection (UAD), aiming to detect anomalies without any labels even when the training set may contain anomalous samples. To achieve FUAD, we pioneer the introduction of Knowledge Distillation (KD) paradigm based on teacher–student framework into the FUAD setting. However, due to the presence of anomalies in the training data, traditional KD methods risk enabling the student to learn the teacher’s representation of anomalies under FUAD setting, thereby resulting in poor anomaly detection performance. To address this issue, we propose a novel Cross-Domain Distillation (CDD) framework based on the widely studied reverse distillation (RD) paradigm. Specifically, we design a Domain-Specific Training, which divides the training set into multiple domains with lower anomaly ratios and train a domain-specific student for each. Cross-Domain Knowledge Aggregation is then performed, where pseudo-normal features generated by domain-specific students collaboratively guide a global student to learn generalized normal representations across all samples. Experimental results on noisy versions of the MVTec AD and VisA datasets demonstrate that our method achieves significant performance improvements over the baseline, validating its effectiveness under FUAD setting.

## 1 Introduction

In the field of industrial image anomaly detection, acquiring or predefining anomalous samples is often impractical. Consequently, Unsupervised Anomaly Detection (UAD), which relies only on normal samples for training, has been extensively studied [5, 7]. To tackle the challenges of UAD task, a variety of methods are proposed, such as those based on memory banks [26, 2], anomaly synthesis [19, 35], and image reconstruction [4, 1]. In recent years, UAD methods based on Knowledge Distillation (KD) have gained increasing attention [6]. Compared to other techniques, they show greater potential in pixel-level anomaly localization. The KD-based UAD methods typically employ a teacher-student framework, which allows the student network to imitate the feature representations of the teacher on normal samples. Since the student is never exposed to anomalous

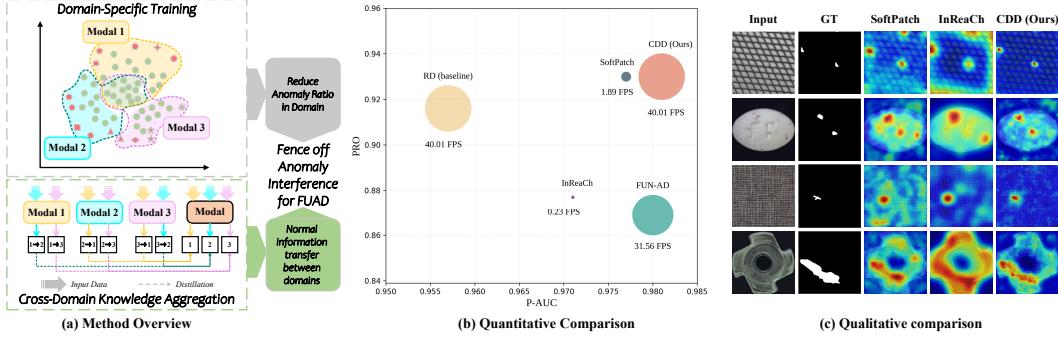


Figure 1: (a) Simplified schematic of Cross-Domain Distillation with 3 domains, where the top represents Domain-Specific Training and the bottom depicts cross-Domain Knowledge Aggregation. (b) Quantitative comparison against other FUAD methods on MVTec AD-noise-0.1 (with 10% anomaly ratio in the train set) [5], with P-AUC (x-axis), PRO (y-axis), and circle size indicating FPS (larger means faster inference), proves that CDD has the best overall performance. (c) Qualitative comparison with other FUAD methods on MVTec AD-noise-0.1.

samples during training, its ability to generate teacher’s anomaly features is limited. This discrepancy in feature mimicking performance becomes a useful signal for anomaly identification.

In real-world scenarios, however, it is often inevitable that the collected data contain a small proportion of anomalous samples. Relying entirely on manual data cleaning incurs high labor costs. This motivates the need for Fully Unsupervised Anomaly Detection (FUAD), a more practical and challenging setting where the training set may contain unlabeled anomalous samples. Although several studies have begun to explore this task, most existing methods rely heavily on memory banks [17, 24, 16], which introduces additional storage overhead. In contrast, Knowledge Distillation paradigm offers a storage-efficient alternative, yet its potential in FUAD has not been fully explored.

We make two key observations regarding the data under the FUAD setting: (1) From a probabilistic perspective, normal pixels still dominate the training set despite the presence of noise; (2) In terms of feature distribution, teacher representations of normal samples tend to be more compact and stable, making them easier for the student network to learn, whereas anomaly features are more dispersed and less likely to be captured. These insights suggest that, even under the FUAD setting, the student network trained via KD inherently focuses on learning normal representations, leading to poor fitting in anomalous regions. This makes the discrepancy between teacher and student features a reliable signal for anomaly detection, which indicates the feasibility and potential of applying the KD paradigm to the FUAD setting.

However, meanwhile, Knowledge Distillation faces a long-standing over-generalization problem [28, 32, 27]. Even though the student network is trained to learn teacher’s representations only on normal pixels, its learned representation ability may still generalize to anomalous ones, which leads to miss detections when testing on anomalous samples. This issue becomes more pronounced under the FUAD setting. If a certain type of anomaly appears frequently in the training data, the student may learn its common feature patterns and become capable of generating teacher-like representations for similar anomalies during inference.

To address the above challenge, we propose a novel cross-domain distillation framework for FUAD, built upon the widely studied KD-based UAD method Reverse Distillation (RD) [9] with an encoder-decoder architecture. First, our intuition is that *reducing the probability of anomalous samples being learned during training mitigates the student’s tendency to overfit to teacher’s anomaly features*. To achieve this, we design a domain division mechanism that distributes high-confidence normal samples across multiple domains while dispersing potentially anomalous samples, thereby lowering the anomaly ratio within each domain without discarding any data. Considering that data distributions vary across domains and that RD’s student decoder generates anomaly-free features for unseen anomalous samples, we hypothesize that *domain-specific students trained on different domains produce pseudo-normal features when applied to other domains*. Building on this insight, we introduce a Cross-Domain Distillation (CDD) framework: for each domain, we utilize domain-specific students from other domains to generate pseudo-normal features for its samples, guiding the training of a global student decoder. This global student learns to produce anomaly-free features across all samples, both normal and anomalous. Finally, the distance between the features generated

by the global student decoder and the teacher encoder is used to detect and localize anomalies. Our contributions are summarized as follows:

- We are the first to explore the application of the knowledge distillation paradigm to the Fully Unsupervised Anomaly Detection task.
- We propose Domain-Specific Training (DST) as in Fig. 1 (a), which first performs Confidence-Guided Domain Construction to build data domains with low anomaly probability. Then, each domain is used to train a domain-specific student via Domain-Specific Distillation with Regularization.
- We introduce Cross-Domain Knowledge Aggregation (CDKA), where domain-specific students provide pseudo-normal features for each sample to train a global student that integrates information across all domains as depicted in Fig. 1 (a).
- Experimental verification shows that our CDD is significantly higher than the baseline RD, and has better performance and faster inference speed than the previous FUAD methods.

## 2 Related Work

**Unsupervised Anomaly Detection.** Unsupervised Anomaly Detection (UAD) has been widely studied in recent years due to its ability to operate without requiring anomalous samples during training. Existing methods are broadly categorized into the following types: (1) reconstruction-based generative models [4, 1, 29, 25, 33, 37], which learn to reconstruct only normal samples and identify anomalies based on reconstruction errors during inference; (2) density estimation-based methods [8, 12, 39], which assume that normal samples follow a specific distribution in the feature space and detect deviations from this distribution; (3) synthetic anomaly-based approaches [19, 35, 21, 38], which generate pseudo-anomalies using image transformations, external generators, or diffusion models to enhance the model’s ability to perceive anomalies; and (4) methods that incorporate pre-trained models and memory bank mechanisms [26, 2, 15], comparing the features of test samples with those of normal samples to identify anomalies. In recent years, Knowledge Distillation-based UAD methods [6, 20, 28, 32, 27, 3, 22] using the teacher-student framework have emerged as excellent methods for anomaly localization. These methods learn representations of normal regions and detect anomalies by measuring the discrepancy in features between the teacher and student networks on anomalous regions. To mitigate the student’s over-generalization to anomalies, some studies introduce heterogeneous architectures or reverse information flow, such as Reverse Distillation [9] and its variants [30, 13, 11, 18, 14, 36], which further improve anomaly detection accuracy.

**Fully Unsupervised Anomaly Detection.** Fully Unsupervised Anomaly Detection (FUAD) has attracted increasing attention, owing to its ability to operate without manual annotations and its suitability for tackling noisy training data in real-world scenarios [31]. Existing methods are categorized as follows: (1) SoftPatch [17], based on PatchCore [26], adopts a memory-based patch-level denoising strategy using noise discriminators to mitigate overconfidence. (2) InReaCh [24] builds detection models by associating high-confidence patch channels across training images. (3) FUN-AD [16] leverages nearest-neighbor distances and class homogeneity, employing an iteratively reconstructed memory bank (IRMB) to handle noisy data. However, these methods often rely on explicit memory banks, which impose storage burdens in practice. Knowledge Distillation has shown strong potential in unsupervised anomaly localization without additional storage, but its application to FUAD remains unexplored. This work aims to explore this promising direction.

## 3 Motivation and Assumptions

### 3.1 Rethinking Reverse Distillation for FUAD

**What is Reverse Distillation?** Early KD-based AD methods typically adopt a homogeneous teacher-student framework, where the student only learns the teacher’s representation ability on normal samples. During inference, anomalies are detected by measuring the discrepancy between teacher and student features. Reverse Distillation (RD) [9] builds upon KD by introducing an encoder-decoder structure. The teacher network is a frozen encoder, while the student consists of a trainable one-class bottleneck embedding (OCBE) module  $\mathcal{B}(\cdot; \phi)$  and a trainable decoder  $\mathcal{D}_S(\cdot; \psi)$ .

Let the training set be  $\mathcal{I}_{train}$ . Given a training image  $I_i^{train} \in \mathcal{I}_{train}$ , the teacher extracts multi-layer features  $\mathcal{F}_{T,i} = \mathcal{T}(I_i^{train}) = \{f_{T,i}^l\}_{l=1}^L$ , which are then reconstructed by the student network as  $\mathcal{F}_{S,i} = \mathcal{S}(\mathcal{F}_{T,i}; \theta_S) = \{f_{S,i}^l\}_{l=1}^L$ . The student network is denoted as  $\mathcal{S}(\cdot; \theta_S)$ , with parameters  $\theta_S = \{\phi, \psi\}$ . The training objective is to minimize the cosine distance between teacher and student features across all  $L = 3$  layers on normal samples as:

$$\cos(f_1, f_2) = \frac{f_1 \cdot f_2}{\|f_1\| \|f_2\|} \quad (1)$$

$$\ell_{cos}(\mathcal{F}_T, \mathcal{F}_S) = \sum_{l=1}^L (1 - \cos(f_T^l, f_S^l)) \quad (2)$$

$$\arg \min_{\theta_S} \mathbb{E}_{I_i \sim \mathcal{I}_{train}} \ell_{cos}(\mathcal{F}_{T,i}, \mathcal{S}(\mathcal{F}_{T,i}; \theta_S)) \quad (3)$$

**Why does RD Work for FUAD?** Although Reverse Distillation (RD) is initially designed for training with only normal samples, it demonstrates strong adaptability in Fully Unsupervised Anomaly Detection. We attribute this to two key factors:

(1) **Probability Perspective** - Dominance of Normal Samples

In industrial scenarios, normal samples are much more common than anomalies, which results in low proportion of anomalous images in the training set. Moreover, anomalies typically occupy only a small region within an image. Consequently, the student network, driven by the dominance of normal samples, primarily learns to represent normal features, while the sparsity of anomalies limits their impact on the optimization process.

(2) **Distribution Perspective** - Concentrated Normal vs. Diverse Anomalous

Normal samples exhibit compact and consistent feature patterns, while anomalous samples are diverse and scattered. This makes it difficult for the student to generalize learned anomalous features.

**Challenges of Applying RD to FUAD.** In FUAD task, the training set naturally includes a certain proportion of anomalous samples. If specific anomaly patterns appear repeatedly during training, the student can easily learn to reconstruct the teacher features of these common anomalies. This results in poor discrimination against similar anomalies during testing and further intensifies over-generalization. Therefore, the key challenge in applying RD to FUAD is *how to prevent the student from modeling common anomaly patterns during training, to ensure that it generates anomaly-free features*.

### 3.2 Assumptions

To address over-generalization problem in FUAD, we propose two assumptions based on the diversity and sparsity of anomalies, guiding the following design of our method Cross-Domain Distillation.

**Assumption 1 (Limited Representation of Rare Anomalies)** *When a particular anomaly type is sufficiently rare in training data, the student fails to learn its corresponding teacher anomaly features, and instead tends to produce features that closely resemble normal patterns.*

Due to the consistency of normal samples and the diversity of anomalies (i.e., anomalies exhibit multiple distinct patterns), we assume the training set contains one normal type and  $M_{train}$  anomaly types, expressed as:

$$\mathcal{I}_{train} = \mathcal{N} \cup \mathcal{A} = \mathcal{N} \cup \bigcup_{m=1}^{M_{train}} \mathcal{A}_m \quad (4)$$

where  $\mathcal{N}$  denotes the set of normal samples,  $\mathcal{A}_m$  denotes the set of the  $m$ -th anomaly type, and:

$$\mathbb{P}(\mathcal{N}) \gg \mathbb{P}(\mathcal{A}_m) \quad \forall m = 1, \dots, M_{train} \quad (5)$$

Following Empirical Risk Minimization (ERM), the training objective is to minimize the distance between student features and teacher features over all samples. The empirical risk can be expressed as

$$\mathcal{L} = \mathbb{P}(\mathcal{N}) \cdot \mathbb{E}_{I_i \sim \mathcal{N}} [\ell_{cos}(\mathcal{F}_{T,i}, \mathcal{F}_{S,i})] + \sum_{m=1}^{M_{train}} \mathbb{P}(\mathcal{A}_m) \cdot \mathbb{E}_{I_j \sim \mathcal{A}_m} [\ell_{cos}(\mathcal{F}_{T,i}, \mathcal{F}_{S,i})] \quad (6)$$

The gradient of parameters  $\theta_S$  is:

$$\frac{\partial \mathcal{L}}{\partial \theta_S} = \mathbb{P}(\mathcal{N}) \cdot \mathbb{E}_{I_i \sim \mathcal{N}} \left[ \frac{\partial \ell_{cos}}{\partial \theta_S} \right] + \sum_{m=1}^{M_{train}} \mathbb{P}(\mathcal{A}_m) \cdot \mathbb{E}_{I_j \sim \mathcal{A}_m} \left[ \frac{\partial \ell_{cos}}{\partial \theta_S} \right] \quad (7)$$

If  $\mathbb{P}(\mathcal{A}_m)$  is small enough, the contribution of the anomaly type  $\mathcal{A}_m$  to the gradient is negligible. Thus, the student receives limited learning signals for this type of anomaly and fails to reconstruct the corresponding teacher features effectively.

**Assumption 2 (Lack of Cross-Anomaly Generalization)** *Even if a student learns to reconstruct some specific anomaly patterns during training, this reconstruction ability is not generalized to other unseen anomaly types.*

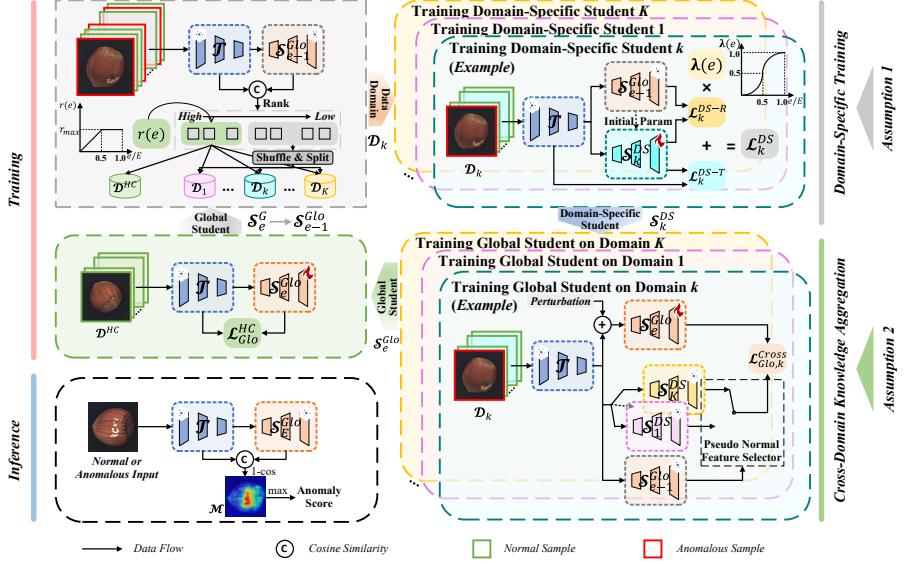


Figure 2: Overall framework of our proposed CDD.

This assumption is based on the diversity of anomalies. Anomalies are often unstructured and come from different sources or physical mechanisms. As a result, they follow multiple, structurally different patterns in the feature space:

$$\mathcal{F}_{T,i} \mid I_i \in \mathcal{A}_m \sim \mathcal{P}_m \quad (8)$$

where each pattern  $\mathcal{P}_m$  represents the teacher's feature distribution for the  $m$ -th type of anomaly. The total number of types is  $M$ , which may even be infinite in practice.

During training, the student network only sees a subset of these anomaly patterns:

$$\mathcal{P}_{\text{train}} = \{\mathcal{P}_1, \dots, \mathcal{P}_{M_{\text{train}}}\}, \quad M_{\text{train}} \ll M \quad (9)$$

According to the **No Free Lunch** theorem, if an input anomalous sample  $I_j \sim \mathcal{P}_{m'}$  with  $\mathcal{P}_{m'} \notin \mathcal{P}_{\text{train}}$ , its distribution is outside the training support. Then, the student may fail to generate the correct teacher features:

$$\mathcal{F}_{S,j} \not\sim \mathcal{F}_{T,j}, \quad I_j \notin \mathcal{P}_{\text{train}} \quad (10)$$

Since normal samples dominate the training set, the student tends to generate features similar to the normal distribution.

## 4 Method

**Problem Definition.** In FUAD, we denote the training set as  $\mathcal{I}_{\text{train}} = \{I_i^{\text{train}}\}_{i=1}^N$ , where each image  $I_i^{\text{train}} \in \{\mathcal{N}, \mathcal{A}\}$  is unlabeled and may be normal or anomalous. The test set  $\mathcal{I}_{\text{test}} = \{I_j^{\text{test}}\}_{j=1}^M$  comprises both normal and anomalous images, with normal samples following the same distribution as  $\mathcal{I}_{\text{train}}$ . The objective is to learn the distribution of normal samples from  $\mathcal{I}_{\text{train}}$  to detect anomalies in  $\mathcal{I}_{\text{test}}$ .

**Overview.** Fig. 2 illustrates the training process of each epoch (top and lower right) and the inference process (lower left). All teacher and student networks follow the design of Reverse Distillation. The teacher is a WideResNet-50 [34] pre-trained on ImageNet [10]. And each student includes an OCBE module and a decoder.

Each training epoch consists of two stages: Domain-Specific Training and Cross-Domain Knowledge Aggregation. In the first stage, we propose Confidence-Guided Domain Construction to extract high-confidence normal samples from the original training set and use them as the intersection between multiple data domains. In this way, each domain has a reduced anomaly ratio compared to the full dataset. Then, we train a domain-specific student for each domain using Domain-Specific Distillation with Regularization. Based on *Assumption 1*, these students ease off from modeling anomaly features and thus focus on modeling normal features in their local domains. The second stage Cross-Domain Knowledge Aggregation mainly explains how to use the domain-specific students obtained in the first stage to train a global student that reconstructs normal features on all samples. According to *Assumption 2*, for anomalous samples in a specific domain  $k$ , domain-specific students that are not trained on domain  $k$  still generates normal-like features. We use these features as pseudo-normal supervision signals to perform Cross-Domain Pseudo-Normal Feature Distillation for the global student. After that, we further distill the global student using the teacher on high-confidence normal samples, enabling it to effectively learn the reliable reconstruction of normal patterns.

The lower left part of Fig. 2 depicts the inference process. During inference, for each image  $I_j^{\text{test}} \in \mathcal{I}_{\text{test}}$ , cosine distances across multi-layer features generated by the teacher  $T$  and the global student trained for  $E$  epochs

$\mathcal{S}_E^{Glo}$  are fused to generate a pixel-level anomaly map  $\mathcal{M}$ , whose maximum value serves as the image-level anomaly score  $s$ :

$$\mathcal{M}(h, w) = \sum_{l=1}^L \left( 1 - \cos(f_T^l(h, w), f_{\mathcal{S}_E^{Glo}}^l(h, w)) \right), s = \max(\mathcal{M}) \quad (11)$$

#### 4.1 Domain-Specific Training

**Confidence-Guided Domain Construction** Based on *Assumption 1*, reducing the anomaly probability in the training set helps the student better learn normal patterns. A naive way to achieve this is to retain only high-confidence normal samples or discard low-confidence anomalous ones. However, such strategies fail to fully utilize the training data, as potentially useful normal regions are also discarded along with the anomalies.

To address this issue, we introduce a confidence-guided strategy on top of naive equal partitioning. Specifically, we inject a portion of highly confident normal samples into each domain based on normality confidence scores, which ensures that: (1) The anomaly ratio in each domain becomes lower than that in the original training set, reducing the interference of anomalous samples on the modeling of normal patterns. (2) The normal distribution in each domain remains more similar to the overall normal distribution, mitigating the negative impact of domain partitioning on student normality modeling.

We use the features output by the global student of the previous epoch  $\mathcal{S}_{e-1}^{Glo}$  as the basis for confidence evaluation. For each training sample  $I_i$ , the average cosine similarity between teacher features  $f_T$  and global student features  $f_{\mathcal{S}_{e-1}^G}$  across  $L$  layers is calculated to obtain the corresponding  $\text{Conf}_i$ :

$$\text{Conf}_i = \sum_{l=1}^L \left\{ \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \cos(f_{T,i}^l(h, w), f_{\mathcal{S}_{e-1}^{Glo},i}^l(h, w)) \right\} \quad (12)$$

All samples are sorted by confidence in descending order. The top  $r(e)$  samples form the high-confidence set  $\mathcal{D}^{HC}$ . The confidence threshold  $r(e)$  increases with training, up to 50%. Let  $e$  be the current epoch and  $E$  the total epochs,  $r(e)$  is calculated as

$$r(e) = \min \left( \frac{e}{E}, 0.5 \right) \quad (13)$$

The remaining low-confidence samples are randomly and evenly divided into  $K$  subsets, denoted  $\mathcal{D}_k^{LC}$ ,  $k = 1, \dots, K$ . By combining  $\mathcal{D}^{HC}$  and  $\mathcal{D}_k^{LC}$ , each domain is expressed as

$$\mathcal{D}_k = \mathcal{D}^{HC} \cup \mathcal{D}_k^{LC}, \quad k = 1, \dots, K. \quad (14)$$

**Domain-Specific Distillation with Regularization** After domain construction, we train a corresponding domain-specific student  $\mathcal{S}_k^{DS}$  for the  $k$ -th domain, who learns to reconstruct the features of samples within its corresponding domain. The initial parameters of each domain-specific student are inherited from the global student of the previous epoch  $\mathcal{S}_{e-1}^{Glo}$ . This training process of  $\mathcal{S}_k^{DS}$  follows the basic framework of Reverse Distillation, which minimizes the cosine distance between the features generated by the student  $\mathcal{F}_{\mathcal{S}_k^{DS}}$  and the features of the teacher  $\mathcal{F}_T$ . In this way, the domain-specific students are able to model the teacher's feature representation ability of data in their local domain.

However, even with controlled anomaly ratios and dispersed common anomalies, the domain-specific student may still learn representations of abnormal samples, especially when a particular type of anomaly is overly represented in the domain. To further tackle this problem, we introduce pseudo-normal features generated by the global student obtained from the previous epoch  $\mathcal{S}_{e-1}^{Glo}$  as the regularization signal. As the global student becomes more and more capable of modeling normal patterns during training, it provides useful guidance to help domain-specific students avoid over-learning anomaly features. The loss  $\mathcal{L}_k^{DS}$  used to train each domain-specific student  $\mathcal{S}_k^{DS}$  combines two terms: the primary distillation loss (from the teacher) and the regularization loss (from the global student), which is expressed as

$$\mathcal{L}_k^{DS} = \mathbb{E}_{I_i \sim \mathcal{D}_k} \underbrace{\left( \ell_{cos}(\mathcal{F}_{T,i}, \mathcal{F}_{\mathcal{S}_k^{DS},i}) + \lambda(e) \cdot \ell_{cos}(\mathcal{F}_{\mathcal{S}_{e-1}^{Glo},i}, \mathcal{F}_{\mathcal{S}_k^{DS},i}) \right)}_{\text{Teacher Guidance } \mathcal{L}_k^{DS-T}} \underbrace{}_{\text{Regularization } \mathcal{L}_k^{DS-R}} \quad (15)$$

where  $\lambda(e)$  is a dynamic increasing coefficient that adjusts the regularization strength over the training epochs. It is controlled using an S-shaped scheduling function with  $p = 4.0$  as

$$\lambda(e) = \frac{(e/E)^p}{(e/E)^p + (1 - e/E)^p} \quad (16)$$

#### 4.2 Cross-Domain Knowledge Aggregation

**Cross-Domain Pseudo-Normal Feature Distillation** Due to the high consistency of normal samples across domains, domain-specific students reconstruct correct normal features on normal samples in all domains.

Table 1: Anomaly detection and localization results I-AUC / P-AUC / PRO under *No Overlap* setting on MVTec AD-noise-0.1 with the best in bold. and the second best underlined.

Category	Unsupervised		Fully Unsupervised			
	RD [9]	URD [23]	SoftPatch [17]	InReaCh [24]	FUN-AD [16]	CDD (Ours)
bottle	0.997 / 0.983 / 0.955	0.992 / 0.984 / 0.961	1.000 / 0.986 / 0.956	1.000 / 0.981 / 0.915	1.000 / 0.992 / 0.960	1.000 / 0.987 / 0.959
cable	0.931 / 0.835 / 0.768	0.955 / 0.881 / 0.824	0.996 / 0.984 / 0.919	0.958 / 0.978 / 0.862	0.952 / 0.920 / 0.740	0.981 / 0.969 / 0.891
capsule	0.939 / 0.980 / 0.956	0.951 / 0.981 / 0.958	0.961 / 0.990 / 0.965	0.446 / 0.914 / 0.657	0.922 / 0.987 / 0.855	0.942 / 0.984 / 0.950
carpet	0.985 / 0.988 / 0.957	0.993 / 0.991 / 0.972	0.989 / 0.992 / 0.959	0.980 / 0.992 / 0.958	1.000 / 0.995 / 0.953	0.989 / 0.989 / 0.960
grid	0.956 / 0.994 / 0.979	1.000 / 0.990 / 0.976	0.965 / 0.991 / 0.963	0.917 / 0.983 / 0.929	0.991 / 0.993 / 0.935	1.000 / 0.992 / 0.976
hazelnut	1.000 / 0.992 / 0.936	0.994 / 0.993 / 0.953	1.000 / 0.994 / 0.942	0.997 / 0.988 / 0.907	0.999 / 0.991 / 0.885	1.000 / 0.993 / 0.945
leather	1.000 / 0.995 / 0.988	1.000 / 0.995 / 0.990	1.000 / 0.994 / 0.988	1.000 / 0.992 / 0.985	1.000 / 0.998 / 0.986	1.000 / 0.991 / 0.971
metal_nut	0.988 / 0.833 / 0.859	0.994 / 0.848 / 0.869	0.998 / 0.886 / 0.838	0.970 / 0.958 / 0.887	0.997 / 0.992 / 0.864	1.000 / 0.962 / 0.870
pill	0.960 / 0.966 / 0.956	0.961 / 0.956 / 0.950	0.953 / 0.977 / 0.945	0.889 / 0.956 / 0.883	0.939 / 0.972 / 0.893	0.971 / 0.978 / 0.958
screw	0.980 / 0.995 / 0.983	0.954 / 0.994 / 0.977	0.952 / 0.994 / 0.975	0.779 / 0.982 / 0.936	0.913 / 0.981 / 0.772	0.934 / 0.992 / 0.974
tile	0.988 / 0.961 / 0.858	1.000 / 0.964 / 0.897	1.000 / 0.959 / 0.878	0.999 / 0.965 / 0.878	0.999 / 0.978 / 0.939	0.997 / 0.955 / 0.879
toothbrush	1.000 / 0.991 / 0.939	1.000 / 0.992 / 0.943	1.000 / 0.986 / 0.915	0.990 / 0.989 / 0.904	0.972 / 0.981 / 0.850	0.997 / 0.987 / 0.916
transistor	0.943 / 0.882 / 0.753	0.948 / 0.901 / 0.812	0.996 / 0.952 / 0.819	0.929 / 0.982 / 0.786	0.962 / 0.975 / 0.520	0.998 / 0.980 / 0.831
wood	0.990 / 0.978 / 0.906	0.994 / 0.983 / 0.924	0.997 / 0.979 / 0.912	0.947 / 0.962 / 0.875	1.000 / 0.977 / 0.960	0.993 / 0.979 / 0.916
zipper	0.924 / 0.976 / 0.941	0.861 / 0.973 / 0.926	0.974 / 0.989 / 0.969	0.952 / 0.937 / 0.796	0.984 / 0.970 / 0.925	0.958 / 0.980 / 0.950
Average	0.972 / 0.957 / 0.916	0.973 / 0.962 / 0.929	<b>0.985 / 0.977 / 0.930</b>	0.917 / 0.971 / 0.877	0.975 / 0.980 / 0.869	<b>0.984 / 0.981 / 0.930</b>

Based on *Assumption 2*, the diversity of anomalies prevents domain-specific students from generalizing to out-of-domain anomaly patterns, even if they learn the reconstruction of anomaly features in local domains. Following this idea, we propose using domain-specific students to generate pseudo-normal features for out-of-domain samples. providing supervision for the training of the global student to generate normal features on all samples. To prevent pseudo-normal feature contamination caused by some domain-specific students learning the ability to reconstruct certain types of teacher anomaly features, we design a Consensus-driven Pseudo-Normal Feature Selection strategy.

Specifically, we select the most "consensual" domain-specific student to generate the normal feature supervision for each sample. The core motivation is that for the same sample, multiple domain-specific students that have not been trained on the sample should generate similar normal features. In the implementation, we achieve pseudo-normal feature selection by eliminating outlier features that are more likely to be abnormal features from the output features of domain-specific students with the help of the global student from the previous epoch  $\mathcal{S}_{e-1}^{Glo}$ .

For a sample  $I_i$  from domain  $\mathcal{D}_k$ , we first extract features  $\mathcal{F}_{\mathcal{S}_h^{PS},i} = \{f_{\mathcal{S}_h^{PS},i}^l\}_{l=1}^L, h = \{1, \dots, K\} \setminus k$  using the domain-specific students from domains  $\mathcal{D}_h, h = \{1, \dots, K\} \setminus k$ , and obtain the reference features  $\mathcal{F}_{\mathcal{S}_{e-1}^{Glo},i} = \{f_{\mathcal{S}_{e-1}^{Glo},i}^l\}_{l=1}^L$  the global student from the previous epoch. We construct an affinity matrix  $\text{Aff}_i \in \mathbb{R}^{(K-1) \times 1}$ , where each element measures the cosine similarity between the flattened features of each student and the global student:

$$\text{Aff}_i(h) = \sum_{l=1}^L \cos(f_{\mathcal{S}_h^{PS},i}^l, f_{\mathcal{S}_{e-1}^{Glo},i}^l), \quad h = \{1, \dots, K\} \setminus k \quad (17)$$

The pseudo-normal feature for the training sample is then selected as the one with the highest similarity:

$$\mathcal{F}_{pseudo,i} = \mathcal{F}_{\mathcal{S}_{h^*}^{PS},i}, \quad h^* = \arg \max_h \text{Aff}_i(h) \quad (18)$$

However, the selected pseudo-normal features may still be contaminated with anomaly features. To prevent the trainable global student from overfitting these pseudo-normal features, we inject Gaussian noise with  $\sigma_{noise} = 0.2$  as feature perturbation into its input:

$$\mathcal{F}_{\mathcal{S}_e^{Glo},i}^* = \mathcal{S}(\mathcal{F}_{T,i} + \delta; \theta_{SGlo}), \quad \delta \sim \mathcal{N}(0, \sigma_{noise}^2) \quad (19)$$

The loss of Cross-Domain Pseudo-Normal Feature Distillation  $\mathcal{L}_{Glo}^{Cross}$  is then defined as:

$$\mathcal{L}_{Glo}^{Cross} = \sum_{k=1}^K \mathbb{E}_{I_i \sim \mathcal{D}_k} \ell_{cos}(\mathcal{F}_{pseudo,i}, \mathcal{F}_{\mathcal{S}_e^{Glo},i}^*) \quad (20)$$

**Confident Distillation for High-Confidence Domain** In addition to pseudo-normal feature guidance, we also leverage the previously defined high-confidence sample set  $\mathcal{D}^{HC}$ , using teacher features as direct supervision to further enhance the global student's ability to model true normal patterns:

$$\mathcal{L}_{Glo}^{HC} = \mathbb{E}_{I_i \sim \mathcal{D}^{HC}} \ell_{cos}(\mathcal{F}_{T,i}, \mathcal{F}_{\mathcal{S}_e^{Glo},i}) \quad (21)$$

## 5 Experiments

### 5.1 Experimental Setup

**Datasets.** We conduct experiments on two widely-used datasets: MVTec AD and VisA. Since both datasets are originally designed for unsupervised anomaly detection, we adapt them to the FUAD setting following

Table 2: Anomaly detection and localization results I-AUC / P-AUC / PRO under *Overlap* setting on MVTec AD-noise-0.1 with the best in bold. and the second best underlined.

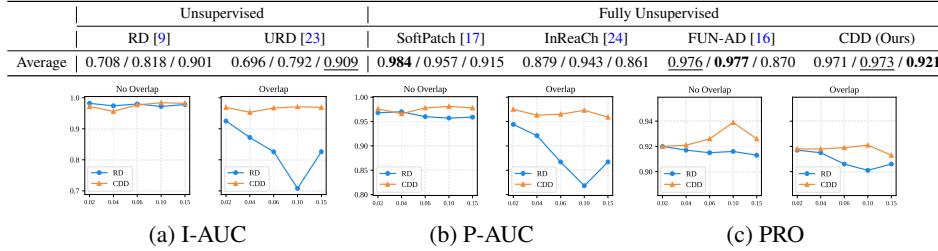


Figure 3: Comparison of anomaly detection performance with baseline RD under different  $R_{\text{noise}}$ .

SoftPatch [17]. Specifically, we keep the normal training images unchanged and randomly inject a portion of anomalous test samples into the training set with a predefined anomaly ratio  $R_{\text{noise}}$ . We evaluate under two settings: (1) *No overlap* setting, where injected anomalous samples are removed from the test set; and (2) *Overlap* setting, where these anomalies remain in the test set, making the task more challenging.

**Implementation Details.** We train a separate model for each category. The backbone follows RD, using a WideResNet-50 pretrained on ImageNet. Following SoftPatch [17], all images are resized to  $256 \times 256$  and then center cropped to  $224 \times 224$  during both training and inference. All domain-specific students and the global student are optimized by its own Adam optimizer with a learning rate of 0.005 and trained for 200 epochs. To smooth the obtained anomaly maps, we apply Gaussian filtering with  $\sigma = 4$ . All our experiments are performed on a single Nvidia GTX 3090 GPU.

**Evaluation Metrics.** We use the area under the ROC curve (AUC) at both image and pixel levels, denoted as I-AUC and P-AUC, to evaluate anomaly detection and localization performance. The per-region-overlap (PRO) metric is also reported to better evaluate the localization performance of anomalies with small sizes.

## 5.2 Anomaly Detection under FUAD setting

**Results on MVTec AD.** We evaluate our proposed Cross-Domain Distillation (CDD) on the MVTec-AD dataset with  $R_{\text{noise}} = 0.1$ , denoted as MVTec-AD-noise-0.1. CDD is compared with unsupervised KD-based UAD methods including RD [9], and FUAD methods, such as SoftPatch [17], InReaCh [24], and FUN-AD [16]. Tab. 1 and Tab. 2 present the anomaly detection and localization results under *No Overlap* and *Overlap* settings, respectively, where each method reports I-AUC, P-AUC, and PRO metrics, all reproduced through 200 epochs of model training under a unified dataset split. In the *No Overlap* setting, CDD matches SoftPatch’s I-AUC while achieving a P-AUC of 0.981 and PRO of 0.930, surpassing all methods including RD in pixel-level localization. In the *Overlap* setting, despite some methods’ performance dropping sharply, CDD retains robustness with a P-AUC of 0.973 and PRO of 0.921, significantly outperforming the baseline and demonstrating strong resistance to anomaly noise. Furthermore, we compare RD and CDD on MVTec AD-noise-{0.2-0.15} as in Fig. 3. At low  $R_{\text{noise}}$ , CDD’s advantage over RD is subtle, but as  $R_{\text{noise}}$  rises, RD becomes unstable, especially in the *Overlap* setting, while CDD shows consistent performance with minimal fluctuations.

**Results on VisA.** For the VisA dataset, we set  $R_{\text{noise}} = 0.05$  (VisA-noise-0.05) based on the ratio of normal to anomalous samples in the original dataset and conduct relevant experiments as in Tab. 3. The compared methods include unsupervised and fully unsupervised AD methods. Our method achieves the best performance in both *No Overlap* and *Overlap* settings. Notably, in the *Overlap* setting, we outperform the baseline RD by 28.0% in I-AUC, 6.8% in P-AUC, and 1.9% in PRO, respectively, demonstrating that our cross-domain training strategy effectively enhances the baseline’s resilience to anomaly interference.

Table 3: Anomaly detection and localization results on VisA-noise-0.05.

Setting	Metrics	RD [9]	SoftPatch [17]	InReaCh [24]	CDD (Ours)
No Overlap	I-AUC	0.945	0.927	0.827	<b>0.954</b>
	P-AUC	0.979	<b>0.985</b>	0.974	<u>0.982</u>
	PRO	0.897	0.904	0.793	<b>0.911</b>
Overlap	I-AUC	0.656	0.924	0.725	<b>0.936</b>
	P-AUC	0.909	<b>0.954</b>	0.914	<b>0.977</b>
	PRO	0.892	0.883	0.721	<b>0.911</b>

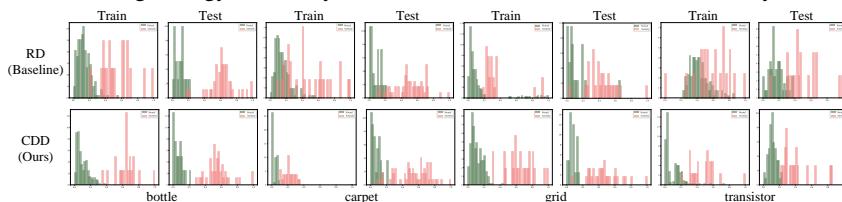


Figure 4: Comparison of histograms of anomaly scores obtained by RD and our CDD.

**Visualization Comparisons.** We perform additional visualization experiments to compare our proposed CDD with the baseline RD. First, we obtain anomaly scores on both the training and test sets of MVTec-

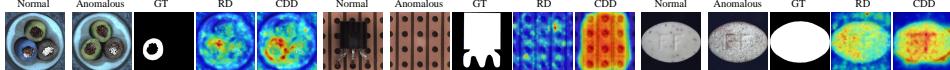


Figure 5: Visualization comparison of anomaly maps generated by RD and our CDD.

AD-noise-0.1 using the trained RD and CDD, generating histograms of anomaly scores for all the samples as depicted in Fig. 4. On one hand, RD proves effective in the FUAD setting, yet it inadvertently learns certain anomaly patterns from the training set, impairing its ability to accurately detect anomalies. Notably, our CDD overcomes this limitation, markedly improving anomaly detection ability on the training set. Fig. 5 further compares anomaly maps generated by RD and CDD. Compared to RD, CDD exhibits greater sensitivity to anomalies, intuitively demonstrating its ability to mitigate overfitting to some extent, preventing the student network from excessively learning the teacher’s anomaly representations.

### 5.3 Ablation Analysis

#### Effectiveness of Proposed Designs.

We first conduct the stepwise ablation experiments to evaluate the effectiveness of module designs based on  $K = 2$  as in Tab. 4. Without any additional designs, the setup reverts to the baseline RD. For Domain-Specific Training (DST), *D.C.* represents a simple even-split domain construction, while *C.G.* integrates Confidence-Guided Domain Construction for improved division. Initially, domain-specific training relies exclusively on the teacher for supervision, with *Reg.* introducing regularization by distilling from the previous global student. For Cross-Domain Knowledge Aggregation (CDKA), *P.N.* denotes the basic cross-domain distillation, generating pseudo-normal features across domains for feature distillation. The inclusion of *F.P.* involves applying feature perturbation to the global student’s input during training. Lastly, *Conf.D.* refers to training the global student directly on High-Confidence Domains to learn teacher representations. The results in Tab. 4 confirm that the addition of each module consistently enhances performance over the baseline.

Table 5: Ablation study of domain number  $K$  on MVTec-noise-0.1.

$K$	I-AUC	P-AUC	PRO	Average
2	0.9836	<b>0.9818</b>	0.9287	0.9647
3	0.9821	0.9793	0.9252	0.9622
4	0.9791	0.9811	0.9271	0.9624
{2,3,3,2}	<b>0.9840</b>	0.9812	<b>0.9297</b>	<b>0.9650</b>
{2,3,4,3,2}	0.9837	0.9806	0.9260	0.9634

Table 4: Ablation study of module effectiveness on MVTec AD-noise-0.1 with  $K = 2$ .

DST	CDKA			I-AUC	P-AUC	PRO	Average
	<i>D.C.</i>	<i>Conf.G.</i>	<i>Reg.</i>	<i>P.N.</i>	<i>F.P.</i>	<i>Conf.D.</i>	
-	-	-	-	-	-	-	0.9721 0.9566 0.9156 0.9481
✓	-	-	✓	-	-	-	0.9701 0.9731 0.9225 0.9552
✓	✓	-	✓	-	-	-	0.9709 0.9764 0.9223 0.9565
✓	✓	-	✓	-	✓	✓	0.9761 0.9779 0.9230 0.9590
✓	✓	-	✓	✓	✓	✓	0.9802 0.9821 0.9287 0.9637
✓	✓	✓	✓	✓	✓	✓	0.9836 0.9818 0.9287 0.9647

Table 6: Ablation study of pseudo-normal feature selection strategies on MVTec-noise-0.1.

$K = 3$					
Select Strategy	I-AUC	P-AUC	PRO	Average	
All	0.9753	0.9747	0.9251	0.9584	
One	Next <i>Consensual</i>	0.9510 <b>0.9821</b>	0.9692 <b>0.9793</b>	0.9142 <b>0.9252</b>	0.9448 <b>0.9622</b>

**Number of Domains.** To investigate the impact of the number of domains  $K$ , we conduct an ablation study on MVTec-AD, with performance results in Tab. 5. Moreover, we observe that as training progresses, the student can gradually generate normal teacher features. In this case, appropriately increasing  $K$  better isolate anomalies. In the later stages, as the global student learns to generate normal features even in anomaly regions, finer domain division becomes less critical, allowing  $K$  to be reduced. To test this, we experimented with dynamic  $K$  strategies. Results show that the {2, 3, 3, 2} strategy achieves a PRO of 0.9297, a 1% improvement over the fixed  $K = 2$ . This indicates that dynamically adjusting  $K$  effectively balances anomaly suppression and normal feature modeling. Therefore, our final design adopts  $K$  varying as {2, 3, 3, 2} across epochs.

**Selection of Pseudo-Normal Features.** We conduct an ablation study on pseudo-normal feature selection strategies, all performed with  $K = 3$ , with results presented in Tab. 6. One strategy, labeled *All*, uses pseudo-normal features generated by domain-specific students from all other domains for distillation. Alternatively, we select features from only one domain, either via our Consensus-driven Pseudo-Normal Feature Selection (denoted as *Consensual*) or by choosing the next domain’s feature (denoted as *Next*, akin to random selection). Results show that our *Consensual* strategy markedly achieves the best performance, which demonstrates that the Consensus-driven strategy significantly enhances cross-domain distillation quality.

## 6 Conclusions

In this paper, we propose a novel Cross-Domain Distillation framework to address the FUAD task. To reduce the impact of anomalies during training, we introduce two key strategies: Domain-Specific Training, which constructs multiple low-anomaly domains and trains corresponding domain-specific students; and Cross-Domain Knowledge Aggregation, which transfers pseudo-normal features in a cross-domain manner to guide a global student. Compared with the original Reverse Distillation (RD) baseline, our approach significantly improves

robustness and accuracy under noisy training conditions. Compared with the original RD baseline, CDD is less affected by anomaly interference under the FUAD setting, as supported by our experimental results.

**Discussion.** Although CDD is implemented based on the RD paradigm, the core design is conceptually general and could be extended to other UAD paradigms. However, our experiments are restricted to RD-based architectures. Future work will focus on adapting and validating CDD under other paradigms, such as feature reconstruction, to further demonstrate its generality.

## References

- [1] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon. 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*, pages 622–637. Springer, 2019. [1](#), [3](#)
- [2] J. Bae, J.-H. Lee, and S. Kim. Pni: industrial anomaly detection using position and neighborhood information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6373–6383, 2023. [1](#), [3](#)
- [3] 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*, pages 128–138, 2024. [3](#)
- [4] 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. [1](#), [3](#)
- [5] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger. Mvtac ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. [1](#), [2](#), [13](#)
- [6] 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*, pages 4183–4192, 2020. [1](#), [3](#)
- [7] 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*, 130(4):947–969, 2022. [1](#)
- [8] 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*, pages 475–489. Springer, 2021. [3](#)
- [9] 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*, pages 9737–9746, 2022. [2](#), [3](#), [7](#), [8](#), [19](#), [20](#), [21](#), [22](#), [23](#), [24](#)
- [10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. [5](#)
- [11] Z. Gu, L. Liu, X. Chen, R. Yi, J. Zhang, Y. Wang, C. Wang, A. Shu, G. Jiang, and L. Ma. Remembering normality: Memory-guided knowledge distillation for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16401–16409, 2023. [3](#)
- [12] D. Gudovskiy, S. Ishizaka, and K. Kozuka. 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*, pages 98–107, 2022. [3](#)
- [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*, pages 6447–6458, 2023. [3](#)
- [14] J. Guo, L. Jia, W. Zhang, H. Li, et al. Reconstrast: Domain-specific anomaly detection via contrastive reconstruction. *Advances in Neural Information Processing Systems*, 36, 2024. [3](#), [19](#), [20](#)
- [15] J. Hyun, S. Kim, G. Jeon, S. H. Kim, K. Bae, and B. J. Kang. Reconpatch: Contrastive patch representation learning for industrial anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2052–2061, 2024. [3](#)

- [16] J. Im, Y. Son, and J. H. Hong. Fun-ad: Fully unsupervised learning for anomaly detection with noisy training data. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9447–9456. IEEE, 2025. [2](#), [3](#), [7](#), [8](#), [19](#), [20](#), [21](#)
- [17] X. Jiang, J. Liu, J. Wang, Q. Nie, K. Wu, Y. Liu, C. Wang, and F. Zheng. Softpatch: Unsupervised anomaly detection with noisy data. *Advances in Neural Information Processing Systems*, 35:15433–15445, 2022. [2](#), [3](#), [7](#), [8](#), [19](#), [20](#), [21](#), [22](#), [23](#), [24](#)
- [18] Y. Jiang, Y. Cao, and W. Shen. A masked reverse knowledge distillation method incorporating global and local information for image anomaly detection. *Knowledge-Based Systems*, 280:110982, 2023. [3](#)
- [19] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9664–9674, 2021. [1](#), [3](#)
- [20] H. Li, Z. Chen, Y. Xu, and J. Hu. Hyperbolic anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17511–17520, 2024. [3](#)
- [21] J. Lin and Y. Yan. A comprehensive augmentation framework for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8742–8749, 2024. [3](#)
- [22] X. Liu, J. Wang, B. Leng, and S. Zhang. Dual-modeling decouple distillation for unsupervised anomaly detection. In *ACM Multimedia 2024*, 2024. URL <https://openreview.net/forum?id=TOMVFf5L6Q>. [3](#)
- [23] X. Liu, J. Wang, B. Leng, and S. Zhang. Unlocking the potential of reverse distillation for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5640–5648, 2025. [7](#), [8](#), [19](#), [20](#)
- [24] D. McIntosh and A. B. Albu. Inter-realization channels: Unsupervised anomaly detection beyond one-class classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6285–6295, 2023. [2](#), [3](#), [7](#), [8](#), [19](#), [20](#), [21](#), [22](#), [23](#)
- [25] P. Perera, R. Nallapati, and B. Xiang. Ogan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2898–2906, 2019. [3](#)
- [26] 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*, pages 14318–14328, 2022. [1](#), [3](#)
- [27] 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*, pages 2592–2602, 2023. [2](#), [3](#)
- [28] 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*, pages 14902–14912, 2021. [2](#), [3](#)
- [29] 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*, 54:30–44, 2019. [3](#)
- [30] 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*, pages 24511–24520, 2023. [3](#)
- [31] C. Wang, W. Zhu, B.-B. Gao, Z. Gan, J. Zhang, Z. Gu, S. Qian, M. Chen, and L. Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22883–22892, 2024. [3](#)
- [32] G. Wang, S. Han, E. Ding, and D. Huang. Student-teacher feature pyramid matching for anomaly detection. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 306. BMVA Press, 2021. [2](#), [3](#)
- [33] Z. You, L. Cui, Y. Shen, K. Yang, X. Lu, Y. Zheng, and X. Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022. [3](#)
- [34] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. [5](#)

- [35] 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*, pages 8330–8339, 2021. [1](#), [3](#)
- [36] 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*, pages 149–158, 2024. [3](#)
- [37] X. Zhang, N. Li, J. Li, T. Dai, Y. Jiang, and S.-T. Xia. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6782–6791, 2023. [3](#)
- [38] X. Zhang, M. Xu, and X. Zhou. 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*, pages 16699–16708, 2024. [3](#)
- [39] Y. Zhou, X. Xu, J. Song, F. Shen, and H. T. Shen. Msflow: Multiscale flow-based framework for unsupervised anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. [3](#)
- [40] 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*, pages 392–408. Springer, 2022. [17](#)

In the **supplementary material**, we organize the content into the following sections:

- Appendix A provides intuitive visualization evidence for the motivation of our method.
- Appendix B presents results on VisA with different noise ratios.
- Appendix C details the algorithm.
- Appendix D includes additional parameter ablation study results.
- Appendix E reports quantitative results for each category on MVTec AD and VisA.
- Appendix F offers more visualization results of the generated anomaly maps.
- Appendix G discusses limitations and proposes future work.

## A Visualizations for Motivation

### A.1 Discussion of the Effectiveness of RD for FUAD

In Sec. 3.1, we discuss why Reverse Distillation (RD) is able to be applied to the Fully Unsupervised Anomaly Detection (FUAD) task, including the Probability Perspective (normal samples dominate) and the Distribution Perspective (normal sample distribution is concentrated, while anomalous distribution is dispersed).

Fig. A1 presents t-SNE visualizations of teacher features for three MVTec [5] categories: carpet, grid, and leather. Green points represent features of normal pixels, and red points represent features of anomalous pixels. The visualizations confirm that features of normal pixels are the majority and clustered, while features of anomalous pixels are typically distant from the normal distribution center, appearing as outliers.

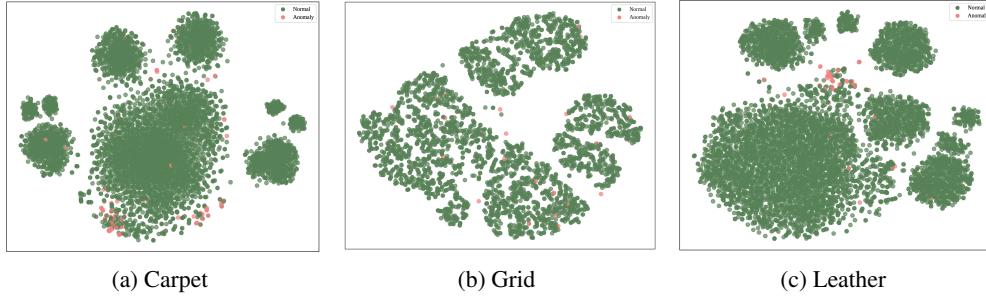


Figure A1: The t-SNE diagrams of feature distribution corresponding to the categories carpet, grid and leather in MVTec AD.

### A.2 Intuitive Evidence of Cross-Domain Distillation

To intuitively demonstrate the feasibility of our proposed Cross-Domain Distillation for the FUAD task, we conduct toy experiments on three categories from the MVTec dataset: cable, metal nut, and transistor, as shown in Fig. A2, Fig. A3, and Fig. A4.

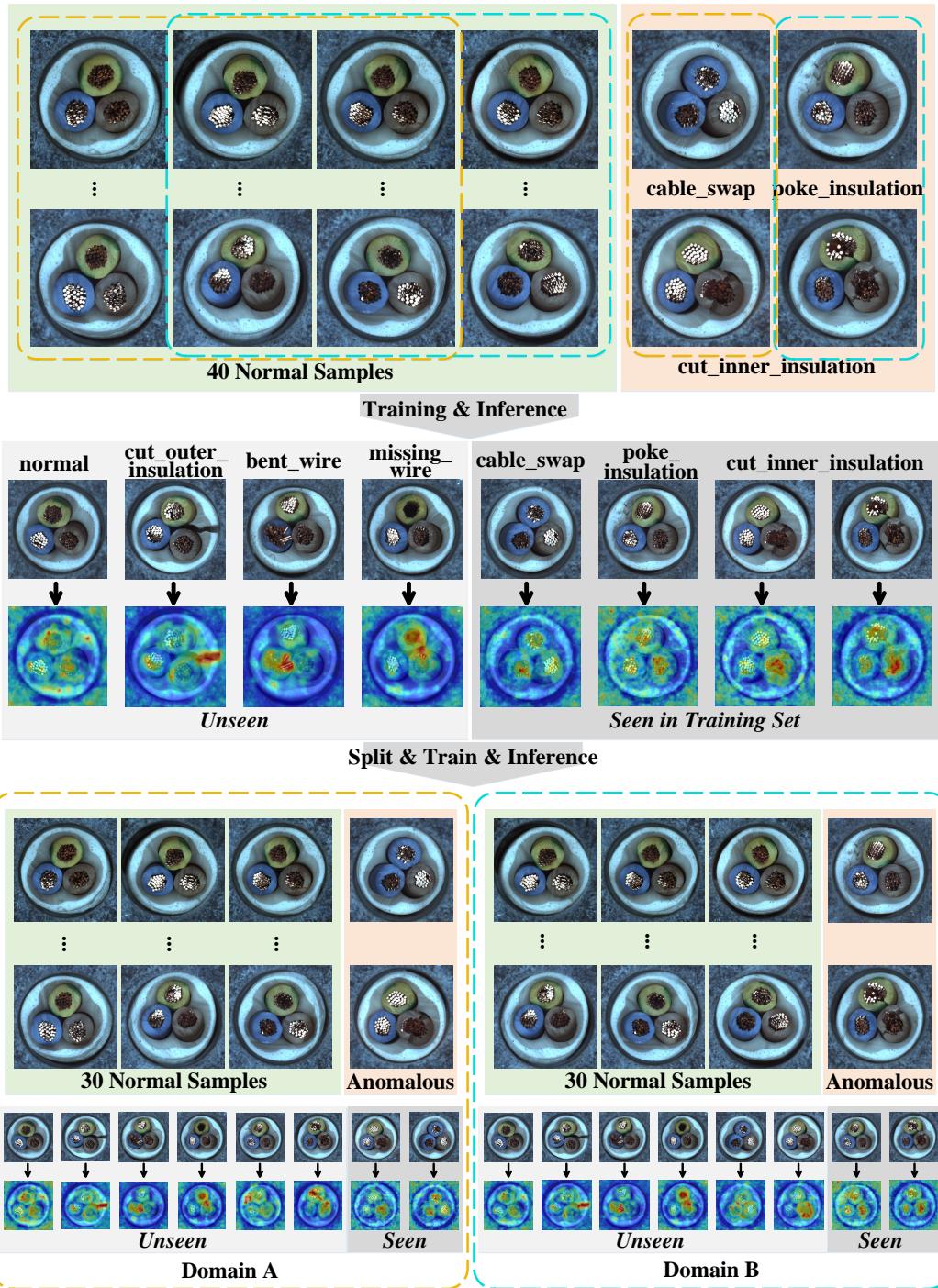


Figure A2: Results of the toy experiment for the category *cable*.

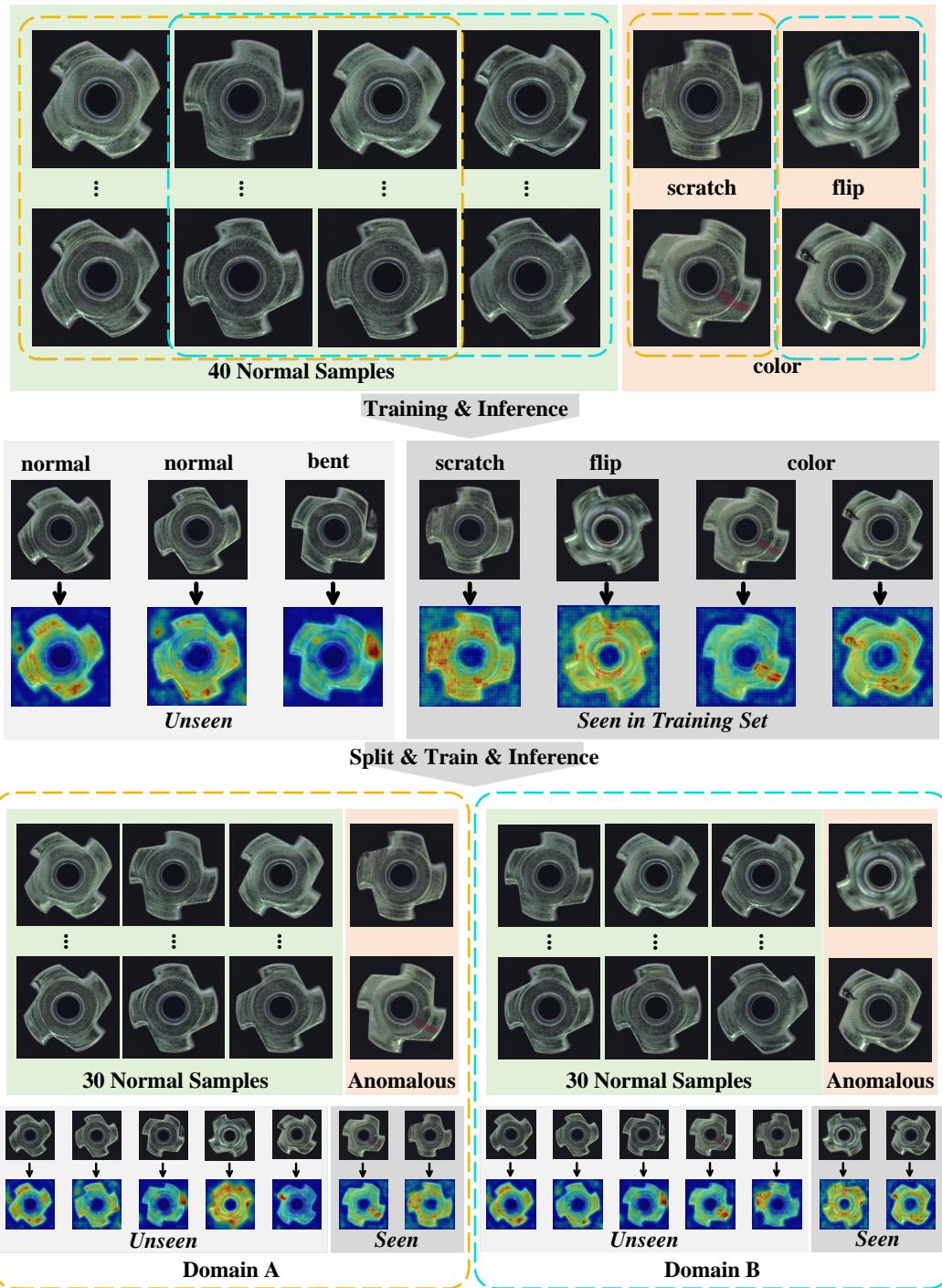


Figure A3: Results of the toy experiment for the category *metal nut*.

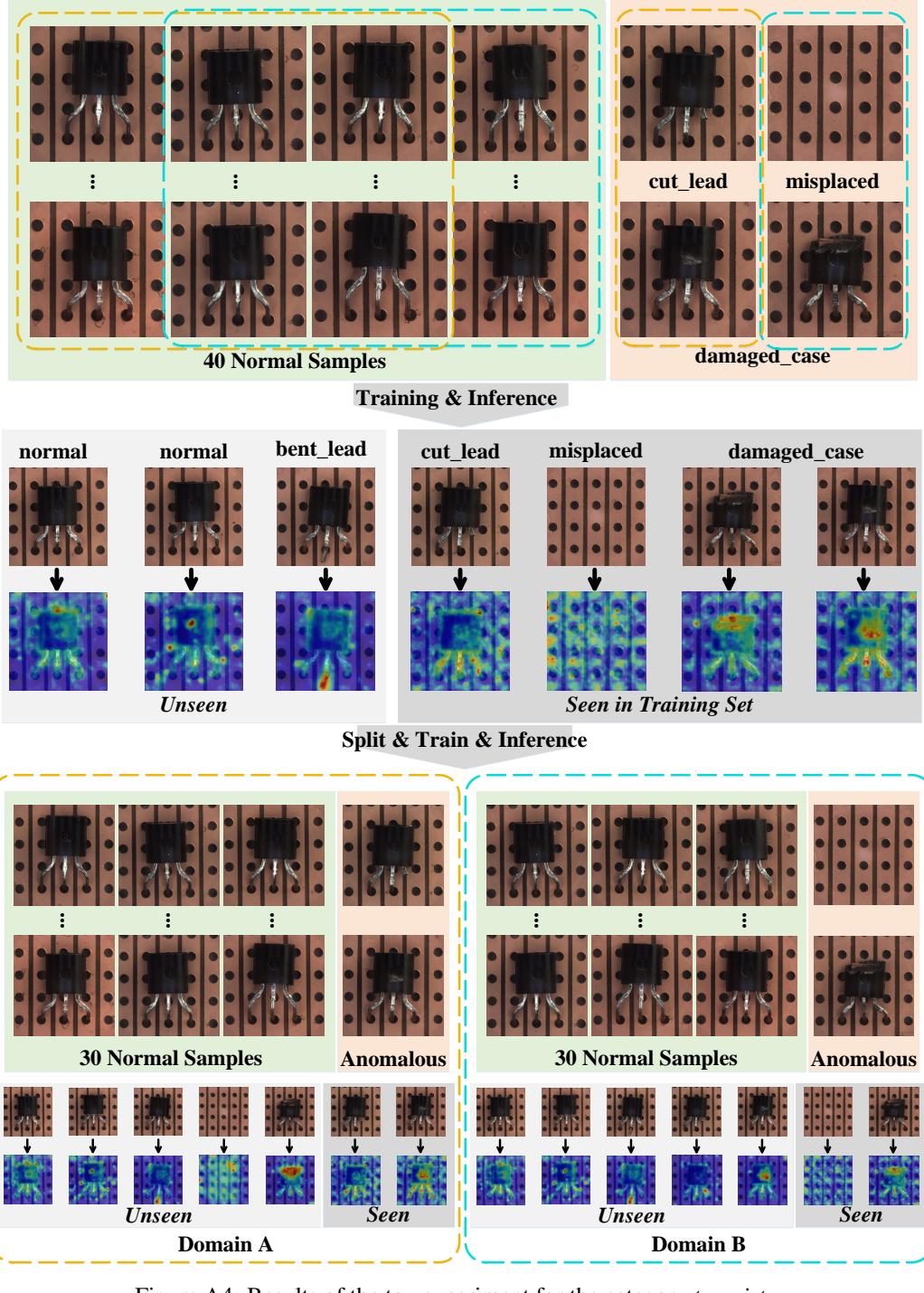


Figure A4: Results of the toy experiment for the category *transistor*.

The process consists of two steps:

- (1) We create a toy train set with a noise ratio of 0.1 by sampling 40 normal samples and 4 anomalous samples from each category, following naive RD (training the student network to learn teacher features from all samples). We test with normal and anomalous samples not used for training and also evaluate samples from the toy training set.

The results show that, compared with the anomalous samples included in the training, better anomaly maps are obtained for anomalous samples not included in the training set. This proves the applicability of RD for

FUAD tasks to a certain extent. However, for some samples that have appeared during training, since the student network has learned the anomaly features of the teacher, there is a high probability that the correct anomaly maps cannot be obtained, which also corresponds to the challenge of applying RD to FUAD tasks mentioned in Sec. 3.1.

(2) Next, we divide the original 44 training samples into two domains. For normal samples, 50% samples are used as an overlapping portion across the two domains. For anomalous samples, we split them evenly.

The results indicate that, even for out-of-domain data matching in-domain anomalous patterns, such as “cut\_inner\_insulation” in cable, “color” in metal nut, and “damaged\_case” in transistor, our method generates correct anomaly maps, which validate our assumptions and support the fundamental motivation of Cross-Domain Distillation.

Additionally, since we use anomaly maps for visualization validation, we believe these results naturally provide a theoretical basis for extending our method to other anomaly detection paradigms.

## B Results on VisA with Different Noise Ratio

Fig. A5 presents a comparison of the anomaly detection and localization results between our method CDD and the baseline RD as the noise ratio in the VisA dataset [40] varies from 0.3 to 0.7. Although under *No Overlap* setting, CDD does not significantly outperform RD in I-AUC and P-AUC metrics, it clearly surpasses RD in the anomaly localization metric PRO. Moreover, as the ratio of anomaly noise increases, the advantage of our method under *Overlap* setting becomes increasingly evident.

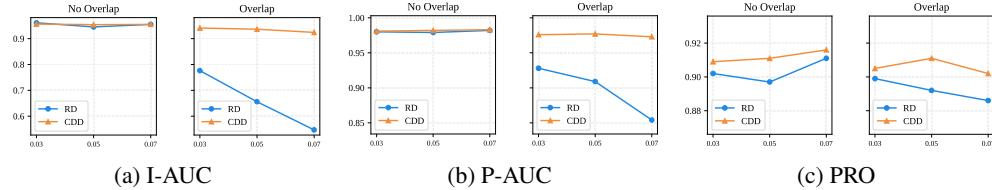


Figure A5: Comparison of anomaly detection performance with baseline RD under different  $R_{\text{noise}}$  on VisA.

## C Algorithm

Algorithm 1 illustrates the complete training process of our proposed Cross-Domain Distillation (CDD). The input is the training set  $\mathcal{I}_{\text{train}}$ , with parameters including the total number of training epochs  $E$ , the number of domains  $K$  (dynamically adjusted during training), the maximum proportion of high-confidence samples in the training set  $r_{\text{normal}}$ , the standard deviation of Gaussian noise for feature perturbation  $\sigma_{\text{noise}}$ , and the parameter  $p$  controlling the smoothness of the regularization term  $\lambda(e)$ . The output is the global student model  $S_E^{\text{Glo}}$  obtained after  $E$  epochs of training.

---

**Algorithm 1** Cross-Domain Distillation (CDD)

---

```

1: Input:  $\mathcal{I}_{train} = \{I_i\}_{i=1}^N, \mathcal{T}$ 
2: Output:  $\mathcal{S}_E^{Glo}$ 
3: Parameters:  $E, K, r_{normal} = 0.5, \sigma_{noise} = 0.2, p = 4.0$ 
4: Initialize global student  $\mathcal{S}_{e-1}^{Glo}$ 
5: for  $e = 0$  to  $E - 1$  do
6:   Compute confidence score  $Conf_i$  for each samples  $I_i \in \mathcal{I}_{train}$  using cosine similarity
      between teacher  $\mathcal{T}$  and previous global student  $\mathcal{S}_{e-1}^{Glo}$ 
7:   Compute threshold  $r(e) = \min(\frac{e}{E}, r_{normal})$ 
8:   Sort samples by confidence, select top  $r(e) \cdot N$  as high-confidence set  $\mathcal{D}^{HC}$ 
9:   Evenly divide the rest samples into  $K$  low-confidence subsets  $\{\mathcal{D}_k^{LC}\}_{k=1}^K$ 
10:  for  $k = 1$  to  $K$  do
11:     $\mathcal{D}_k = \mathcal{D}^{HC} \cup \mathcal{D}_k^{LC}$ 
12:     $\mathcal{S}_k^{DS} \leftarrow \mathcal{S}_{e-1}^{Glo}$ 
13:    Compute the regularization weight  $\lambda(e) = \frac{(e/E)^p}{(e/E)^p + (1-e/E)^p}$ 
14:    for each  $I_i \in \mathcal{D}_k$  do
15:      Train each  $\mathcal{S}_k^{DS}$  on  $I_i$  with the loss  $\mathcal{L}_k^{DS}$  combining teacher guidance and regulariza-
         tion from  $\mathcal{S}_{e-1}^{Glo}$ , where the regularization weight is  $\lambda(e)$ 
16:    end for
17:  end for
18:  for  $k = 1$  to  $K$  do
19:    for each  $I_i \in \mathcal{D}_k$  do
20:      Extract features from other domain-specific students  $\mathcal{F}_{\mathcal{S}_h^{DS}, i}, \forall h \in \{1, \dots, K\} \setminus k$ 
         and features from previous global student  $\mathcal{F}_{\mathcal{S}_{e-1}^{Glo}, i}$ 
21:      Compute affinity to measure feature similarity  $Aff_i(h), \forall h \in \{1, \dots, K\} \setminus k$  using
          $\mathcal{F}_{\mathcal{S}_h^{DS}, i}$  and  $\mathcal{F}_{\mathcal{S}_{e-1}^{Glo}, i}$ 
22:      Select pseudo-normal feature with highest affinity  $\mathcal{F}_{pseudo, i}$ 
23:      Input teacher features with Gaussian noise  $\delta \sim \mathcal{N}(0, \sigma_{noise}^2)$  into global student  $\mathcal{S}_e^{Glo}$ 
         to get  $\mathcal{F}_{\mathcal{S}_e^{Glo}, i}^*$ 
24:      Train  $\mathcal{S}_e^{Glo}$  on sample  $I_i$  by the loss calculated using the cosine distance between
          $\mathcal{F}_{pseudo, i}$  and  $\mathcal{F}_{\mathcal{S}_e^{Glo}, i}^*$ 
25:    end for
26:  end for
27:  for each  $I_i \in \mathcal{D}^{HC}$  do
28:    Train  $\mathcal{S}_e^{Glo}$  on sample  $I_i$  by the loss calculated using teacher guidance
29:  end for
30: end for

```

---

## D More Ablation Studies

### D.1 Ablation Study on Parameter Selection for Confidence Threshold

In Sec. 4.1, we provide the formula for the confidence threshold  $r(e)$ , fixing its maximum proportion at 0.5. This choice is based on the fact that anomalous samples in real-world scenarios are typically below 50%. Setting the maximum threshold at 0.5 balances sample completeness and prevents anomaly noise from contaminating the high-confidence domain. This section validates this choice through experimental evidence, beyond just intuition.

Tab. A1 presents the experimental results for selecting  $r_{normal}$ . During training,  $r(e)$  is computed as  $r(e) = \min(\frac{e}{E}, r_{normal})$ . All experiments are conducted with  $K = 2$ , incorporating training strategies *D.C.*, *P.N.*, *Conf.D.* (refer to Sec. 5.3), and the domain construction strategy *Conf.G.* (refer to Sec. 5.3), which are closely tied to the confidence threshold. No other training techniques were used to avoid interference with the selection of  $r_{normal}$ .

We find that a smaller  $r_{normal}$  yields little improvement in performance. However, when 50% of high-confidence samples are included as normal samples in each domain, anomaly localization improves significantly. Increasing  $r_{normal}$  further, such as to 0.7, slightly enhances anomaly detection but with minimal overall gains. Moreover,

considering that there may be many abnormal scenarios in reality, and the higher the  $r_{\text{normal}}$ , the greater the consumption of training resources, we conclude that  $r_{\text{normal}} = 0.5$  offers a balanced and effective choice.

Table A1: Ablation study of  $r_{\text{normal}}$  on MVTec-noise-0.1 under *No Overlap* setting.

$r_{\text{normal}}$	I-AUC	P-AUC	PRO	Average
0.0	0.9701	0.9731	0.9225	0.9552
0.1	0.9703	0.9732	0.9187	0.9541
0.3	0.9767	0.9727	0.9187	0.9560
0.5	0.9763	0.9795	0.9224	0.9594
0.7	0.9799	0.9794	0.9201	0.9598

## D.2 Ablation Study on Regularization Hyperparameter

During domain-specific distillation to train domain-specific students, we incorporate guidance from the previous global student as a regularization term. The hyperparameter  $\lambda(e)$  for this regularization is designed as an S-shape function in Sec. 4.1. To explain this design choice, we present corresponding experimental results in Tab. A2. We compare  $\lambda(e) = 0.0$ ,  $\lambda(e) = 1.0$ , and a linear design  $\lambda(e) = e/E$ . The results show that the S-shape function achieves the best performance, while fixed or linear weights may introduce noise to the learning of domain-specific students from the global student.

Table A2: Ablation study of regularization hyperparameter  $\lambda(e)$  on MVTec-noise-0.1 under *No Overlap* setting.

$\lambda(e)$	I-AUC	P-AUC	PRO	Average
0.0	0.9802	0.9821	0.9287	0.9637
1.0	0.9768	0.9741	0.9241	0.9583
$e/E$	0.9659	0.9763	0.9225	0.9549
S-shape Function	0.9836	0.9818	0.9289	0.9648

## E Complete Quantitative Results

### E.1 Quantitative Results for Each Category on MVTec AD

Tab. A3, Tab. A4 and Tab. A5 present the image-level AUC (I-AUC), pixel-level AUC (P-AUC), and localization metric PRO for each category on the MVTec AD-noise-0.1 dataset under *No Overlap* and *Overlap* settings.

Table A3: Anomaly detection results I-AUC under *No Overlap* and *Overlap* settings on MVTec AD-noise-0.1 with the best in bold, and the second best underlined.

Setting	No Overlap						Overlap							
	Method	RD [9]	ReContrast [4]	URD [23]	SorPatch [7]	InReCh [24]	FUN-AD [16]	CDD (Ours)	RD [9]	ReContrast [4]	URD [23]	SorPatch [7]	InReCh [24]	FUN-AD [16]
bottle	0.997	0.994	0.992	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	0.742	0.687	0.680	<b>1.000</b>	0.999	<b>1.000</b>	<b>1.000</b>
cable	0.931	0.977	0.955	<u>0.996</u>	0.958	0.952	<u>0.981</u>	0.709	0.743	0.727	<b>0.995</b>	0.871	0.963	<u>0.970</u>
capsule	0.939	<u>0.959</u>	0.951	<b>0.961</b>	0.446	0.922	0.942	0.788	0.835	0.785	0.957	0.456	<u>0.925</u>	<b>0.939</b>
carpet	0.985	<b>1.000</b>	0.993	0.989	0.980	<b>1.000</b>	0.989	0.675	<b>1.000</b>	0.681	0.992	0.978	<b>1.000</b>	0.990
grid	0.956	<b>1.000</b>	<b>1.000</b>	0.965	0.917	0.991	<b>1.000</b>	0.947	0.998	0.817	0.971	0.928	<u>0.995</u>	<b>1.000</b>
hazelnut	<b>1.000</b>	0.999	0.994	<b>1.000</b>	0.997	0.999	<b>1.000</b>	0.443	0.450	0.454	<b>1.000</b>	0.951	<b>1.000</b>	0.986
leather	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	0.762	0.876	0.750	<b>1.000</b>	0.993	<b>1.000</b>	<b>1.000</b>
metal_nut	0.988	0.999	0.994	<u>0.998</u>	0.970	0.997	<b>1.000</b>	0.755	0.763	0.759	<b>0.999</b>	0.935	<u>0.998</u>	0.968
pill	0.960	<b>0.991</b>	0.961	0.953	0.889	0.939	<u>0.971</u>	0.783	0.857	0.784	<u>0.959</u>	0.812	0.946	<b>0.963</b>
screw	<b>0.980</b>	0.979	0.954	0.952	0.779	0.913	0.934	0.716	0.768	0.698	<b>0.934</b>	0.678	0.910	0.849
tile	0.988	0.992	<b>1.000</b>	<b>1.000</b>	0.999	0.999	0.997	0.717	0.720	0.726	<u>0.990</u>	0.982	<b>0.996</b>	0.970
toothbrush	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	0.990	0.972	0.997	0.803	0.800	0.800	<b>1.000</b>	0.964	0.961	<u>0.997</u>
transistor	0.943	0.950	0.948	<u>0.996</u>	0.929	0.962	<b>0.998</b>	0.448	0.451	0.450	<b>0.996</b>	0.823	0.968	<b>0.996</b>
wood	0.990	0.990	0.994	<u>0.997</u>	0.947	<b>1.000</b>	0.993	0.594	0.615	0.630	<u>0.990</u>	0.868	<b>1.000</b>	0.985
zipper	0.924	<u>0.983</u>	0.861	0.974	0.952	<b>0.984</b>	0.958	0.745	0.856	0.699	<u>0.973</u>	0.946	<b>0.983</b>	0.957
Average	0.972	<b>0.988</b>	0.973	<u>0.985</u>	0.917	0.975	0.984	0.708	0.761	0.696	<b>0.984</b>	0.879	<u>0.976</u>	0.971

Table A4: Anomaly localization results P-AUC under *No Overlap* and *Overlap* settings on MVTec AD-noise-0.1 with the best in bold. and the second best underlined.

Setting	No Overlap						Overlap							
	RD [9]	ReContrast [14]	URD [23]	SoftPatch [17]	InReCh [24]	FUN-AD [16]	CDD (Ours)	RD [9]	ReContrast [14]	URD [23]	SoftPatch [17]	InReCh [24]	FUN-AD [16]	CDD (Ours)
bottle	0.983	0.984	0.984	<u>0.988</u>	0.981	<b>0.992</b>	0.987	0.917	0.902	0.854	<u>0.983</u>	0.972	<b>0.991</b>	0.982
cable	0.835	0.924	0.881	<b>0.989</b>	<u>0.978</u>	0.920	0.969	0.727	0.828	0.753	<b>0.981</b>	0.957	0.931	<u>0.962</u>
capsule	0.980	0.976	0.981	<u>0.990</u>	0.914	<u>0.987</u>	0.984	0.889	0.917	0.878	<b>0.989</b>	0.861	<u>0.986</u>	0.978
carpet	0.988	0.990	0.991	<u>0.991</u>	<u>0.992</u>	<b>0.995</b>	0.989	0.732	<u>0.990</u>	0.717	<u>0.990</u>	0.987	<u>0.994</u>	0.988
grid	<b>0.994</b>	0.990	0.990	0.992	0.983	<u>0.993</u>	0.992	<b>0.990</b>	0.984	0.952	0.972	0.975	<u>0.990</u>	0.989
hazelnut	0.992	0.992	<b>0.993</b>	<b>0.993</b>	0.988	0.991	<b>0.993</b>	0.774	0.870	0.768	0.887	0.965	<b>0.992</b>	<u>0.983</u>
leather	0.995	<u>0.996</u>	0.995	0.994	0.992	<b>0.998</b>	0.991	0.848	0.941	0.775	<u>0.993</u>	0.990	<b>0.998</b>	0.990
metal_nut	0.833	0.870	0.848	0.857	0.958	<b>0.992</b>	<u>0.962</u>	0.736	0.796	0.719	0.856	0.925	<b>0.993</b>	<u>0.960</u>
pill	0.966	<b>0.985</b>	0.956	<u>0.981</u>	0.956	0.972	0.978	0.889	0.971	0.838	<u>0.975</u>	0.942	0.971	<b>0.976</b>
screw	<b>0.995</b>	0.994	0.994	0.994	0.982	0.981	0.992	0.865	0.974	0.803	0.961	0.964	<b>0.981</b>	0.974
tile	0.961	<u>0.969</u>	0.964	0.965	0.965	<b>0.978</b>	0.955	0.823	0.834	0.820	<u>0.957</u>	0.955	<b>0.980</b>	0.946
toothbrush	0.991	<b>0.992</b>	<b>0.992</b>	0.987	0.989	0.981	0.987	0.953	0.929	0.952	<b>0.985</b>	0.979	0.984	
transistor	0.882	0.917	0.901	0.972	<b>0.982</b>	0.975	<u>0.980</u>	0.587	0.720	0.526	0.922	0.911	<u>0.953</u>	<b>0.961</b>
wood	0.978	<u>0.982</u>	<b>0.983</b>	0.938	0.962	0.977	0.979	0.658	0.723	0.671	0.917	0.874	<b>0.957</b>	<u>0.945</u>
zipper	0.976	0.977	0.973	<b>0.988</b>	0.937	0.970	<u>0.980</u>	0.888	0.922	0.853	<b>0.986</b>	0.881	0.962	<u>0.971</u>
Average	0.957	0.969	0.962	0.975	0.971	<b>0.980</b>	<u>0.981</u>	0.818	0.887	0.792	0.957	0.943	<b>0.977</b>	0.973

Table A5: Anomaly localization results PRO under *No Overlap* and *Overlap* settings on MVTec AD-noise-0.1 with the best in bold. and the second best underlined.

Setting	No Overlap						Overlap							
	RD [9]	ReContrast [14]	URD [23]	SoftPatch [17]	InReCh [24]	FUN-AD [16]	CDD (Ours)	RD [9]	ReContrast [14]	URD [23]	SoftPatch [17]	InReCh [24]	FUN-AD [16]	CDD (Ours)
bottle	0.955	0.958	<b>0.961</b>	0.956	0.915	<u>0.960</u>	0.959	0.928	0.947	0.947	<u>0.954</u>	0.917	<b>0.963</b>	0.949
cable	0.768	0.847	0.824	<b>0.919</b>	0.862	0.740	<u>0.891</u>	0.740	0.825	0.812	<b>0.918</b>	0.850	0.770	<u>0.888</u>
capsule	0.956	0.953	<u>0.958</u>	<b>0.965</b>	0.657	0.855	0.950	0.953	0.947	<u>0.956</u>	<b>0.959</b>	0.599	0.862	0.940
carpet	0.957	<u>0.968</u>	<b>0.972</b>	0.959	0.958	0.953	0.960	0.932	<b>0.973</b>	0.915	0.966	0.962	0.954	<u>0.967</u>
grid	<b>0.979</b>	0.978	0.976	0.963	0.929	0.935	0.976	0.966	<u>0.975</u>	0.965	0.949	0.914	0.932	<u>0.971</u>
hazelnut	0.936	0.931	<b>0.953</b>	0.942	0.907	0.885	<u>0.945</u>	0.930	<u>0.937</u>	<b>0.948</b>	0.879	0.898	0.883	0.936
leather	0.988	<b>0.991</b>	0.990	0.988	0.985	0.986	0.971	<u>0.988</u>	<u>0.991</u>	0.985	0.988	0.981	0.985	0.970
metal_nut	0.859	0.870	0.869	0.838	<b>0.887</b>	0.864	0.870	0.830	0.877	0.865	0.832	0.869	<b>0.879</b>	0.858
pill	0.956	<b>0.969</b>	0.950	0.945	0.883	0.893	<u>0.958</u>	0.957	<u>0.969</u>	<u>0.951</u>	0.942	0.886	0.892	0.947
screw	<b>0.983</b>	0.977	0.977	0.975	0.936	0.772	0.974	<b>0.978</b>	0.976	0.975	0.923	0.923	0.788	<u>0.976</u>
tile	0.858	0.894	<u>0.897</u>	0.878	0.878	<b>0.939</b>	0.879	0.861	0.883	0.822	0.874	0.863	<b>0.939</b>	<u>0.887</u>
toothbrush	0.939	0.940	<b>0.943</b>	0.915	0.904	0.850	0.916	0.918	<u>0.937</u>	<b>0.938</b>	0.892	0.899	0.829	0.907
transistor	0.753	0.786	0.812	<u>0.819</u>	0.786	0.520	<b>0.831</b>	0.700	0.718	0.738	0.786	0.758	0.505	<b>0.788</b>
wood	0.906	0.924	0.924	0.912	0.875	<b>0.960</b>	0.916	0.888	<u>0.920</u>	0.889	0.890	0.840	<b>0.957</b>	0.889
zipper	0.941	0.938	0.926	<b>0.969</b>	0.796	0.925	<u>0.950</u>	0.942	0.924	0.922	<b>0.967</b>	0.752	0.916	0.938
Average	0.916	0.928	0.929	<b>0.930</b>	0.877	0.869	<b>0.930</b>	0.901	<u>0.920</u>	0.909	0.915	0.861	0.870	<b>0.921</b>

## E.2 Quantitative Results for Each Category on VisA

Tab. A6, Tab. A7 and Tab. A8 present the image-level AUC (I-AUC), pixel-level AUC (P-AUC), and localization metric PRO for each category on the VisA-noise-0.05 dataset under *No Overlap* and *Overlap* settings.

Table A6: Anomaly detection results I-AUC under *No Overlap* and *Overlap* settings on VisA-noise-0.05 with the best in bold. and the second best underlined.

Setting	No Overlap						Overlap						
	RD [9]	SoftPatch [17]	InReCh [24]	FUN-AD [16]	CDD (Ours)	RD [9]	SoftPatch [17]	InReCh [24]	FUN-AD [16]	CDD (Ours)	RD [9]	SoftPatch [17]	
candle	0.928	<b>0.978</b>	0.936	0.931	<u>0.954</u>	0.617	<b>0.977</b>	0.770	0.938	<u>0.948</u>			
capsules	0.800	0.716	0.548	<b>0.886</b>	<u>0.807</u>	0.577	0.695	0.496	<u>0.885</u>	<b>0.885</b>			
casewh	0.962	<u>0.975</u>	0.850	0.945	<b>0.984</b>	0.751	0.970	0.809	0.943	<b>0.983</b>			
chewinggum	0.978	<b>0.987</b>	0.747	<u>0.980</u>	0.977	0.763	<b>0.988</b>	0.716	<u>0.983</u>	0.982			
fryum	<b>0.979</b>	0.944	0.873	0.953	0.966	0.794	0.944	0.837	<u>0.951</u>	<b>0.965</b>			
macaroni1	<u>0.977</u>	0.960	0.861	0.936	<b>0.983</b>	0.696	<b>0.956</b>	0.656	<u>0.950</u>	0.937			
macaroni2	0.810	0.675	0.660	0.775	<b>0.873</b>	0.565	0.680	0.488	0.802	<b>0.825</b>			
pcb1	0.979	0.980	0.645	0.934	<b>0.984</b>	0.539	0.971	0.600	0.942	<b>0.978</b>			
pcb2	<b>0.945</b>	0.934	0.921	0.896	<u>0.944</u>	0.561	<u>0.944</u>	0.889	0.927	<b>0.952</b>			
pcb3	<b>0.988</b>	0.982	0.946	0.888	<u>0.983</u>	0.590	<b>0.971</b>	0.706	0.912	<u>0.946</u>			
pcb4	0.996	<b>0.997</b>	0.980	0.968	<b>0.997</b>	0.647	<u>0.997</u>	0.924	0.974	<b>0.998</b>			
pipe_fryum	<b>0.995</b>	0.992	0.962	<u>0.994</u>	0.992	0.776	<u>0.993</u>	0.809	<b>0.995</b>	0.984			
Average	0.945	0.927	0.827	0.924	<b>0.954</b>	0.656	0.924	0.725	<u>0.934</u>	<b>0.936</b>			

Table A7: Anomaly localization results P-AUC under *No Overlap* and *Overlap* settings on VisA-noise-0.05 with the best in bold. and the second best underlined.

Setting	No Overlap					Overlap				
	RD [9]	SoftPatch [17]	InReaCh [24]	FUN-AD [16]	CDD (Ours)	RD [9]	SoftPatch [17]	InReaCh [24]	FUN-AD [16]	CDD (Ours)
candle	0.989	<b>0.995</b>	0.987	0.990	<u>0.992</u>	0.956	<b>0.993</b>	0.851	0.989	0.986
capsules	<b>0.990</b>	0.985	0.949	<b>0.990</b>	0.987	0.826	0.961	0.887	<b>0.989</b>	0.987
cashew	0.937	0.986	0.983	<u>0.993</u>	0.938	0.931	<u>0.986</u>	0.974	<b>0.993</b>	0.937
chewinggum	0.970	<u>0.989</u>	0.920	<u>0.993</u>	0.982	0.837	<u>0.990</u>	0.914	<b>0.993</b>	0.978
fryum	0.941	0.924	<b>0.959</b>	<u>0.951</u>	0.942	0.908	0.918	<u>0.949</u>	<b>0.951</b>	0.942
macaroni1	0.994	<b>0.998</b>	0.979	0.994	<b>0.998</b>	0.980	<b>0.996</b>	0.857	0.994	<u>0.995</u>
macaroni2	0.986	<u>0.987</u>	0.968	0.982	<b>0.995</b>	0.936	0.743	0.831	<u>0.984</u>	<b>0.986</b>
pcb1	0.994	<b>0.999</b>	<u>0.996</u>	0.995	0.992	0.885	<b>0.998</b>	0.946	<u>0.993</u>	0.986
pcb2	0.983	<b>0.985</b>	0.977	0.953	<b>0.985</b>	0.912	0.968	0.916	0.959	<b>0.978</b>
pcb3	<u>0.995</u>	<b>0.996</b>	0.990	0.982	<u>0.995</u>	0.960	<b>0.992</b>	0.954	0.966	<u>0.990</u>
pcb4	<b>0.987</b>	<b>0.987</b>	0.985	0.986	0.986	0.871	0.917	0.904	<b>0.981</b>	0.979
pipe_fryum	0.985	0.989	<b>0.994</b>	<u>0.993</u>	0.987	0.905	<u>0.989</u>	0.982	<b>0.993</b>	0.984
Average	0.979	<b>0.985</b>	0.974	<u>0.984</u>	0.982	0.909	0.954	0.914	<b>0.982</b>	<u>0.977</u>

Table A8: Anomaly localization results PRO under *No Overlap* and *Overlap* settings on VisA-noise-0.05 with the best in bold. and the second best underlined.

Setting	No Overlap					Overlap				
	RD [9]	SoftPatch [17]	InReaCh [24]	FUN-AD [16]	CDD (Ours)	RD [9]	SoftPatch [17]	InReaCh [24]	FUN-AD [16]	CDD (Ours)
candle	0.911	<b>0.928</b>	0.918	0.801	<b>0.928</b>	0.911	<b>0.946</b>	0.861	0.831	<u>0.939</u>
capsules	<u>0.941</u>	0.831	0.655	0.836	<b>0.950</b>	<u>0.944</u>	0.735	0.607	0.819	<b>0.958</b>
cashew	0.810	<b>0.940</b>	0.521	<u>0.891</u>	0.867	0.790	<b>0.937</b>	0.462	<u>0.875</u>	0.866
chewinggum	0.749	<b>0.883</b>	0.451	<u>0.820</u>	0.819	0.731	<b>0.884</b>	0.426	0.809	<u>0.812</u>
fryum	<b>0.893</b>	0.825	0.797	0.699	<u>0.886</u>	<b>0.899</b>	0.816	0.785	0.695	<u>0.889</u>
macaroni1	0.954	<u>0.968</u>	0.876	0.892	<b>0.972</b>	0.938	<u>0.965</u>	0.798	0.863	<b>0.974</b>
macaroni2	0.903	0.902	0.860	0.800	<b>0.967</b>	<u>0.916</u>	0.800	0.722	0.801	<b>0.971</b>
pcb1	<b>0.945</b>	0.944	0.899	0.778	0.913	<u>0.901</u>	<b>0.919</b>	0.671	0.721	0.887
pcb2	<b>0.881</b>	<u>0.874</u>	0.844	0.658	0.872	<b>0.890</b>	0.873	0.761	0.682	<u>0.875</u>
pcb3	<b>0.946</b>	0.931	0.903	0.842	<u>0.934</u>	<b>0.940</b>	0.923	0.862	0.828	<u>0.936</u>
pcb4	<b>0.879</b>	0.861	0.832	0.646	0.864	<b>0.896</b>	0.846	0.758	0.707	<u>0.875</u>
pipe_fryum	0.956	<u>0.961</u>	<b>0.962</b>	0.895	0.955	<b>0.952</b>	<b>0.952</b>	0.940	0.883	0.947
Average	0.897	<b>0.904</b>	0.793	0.797	<b>0.911</b>	<u>0.892</u>	0.883	0.721	0.793	<b>0.911</b>

## F More Qualitative Results

### F.1 Qualitative Results on MVTec

Figure A6 provides a visualization comparison of anomaly maps generated by FUAD methods, including SoftPatch [17], InReaCh [24], the baseline RD [9], and our proposed CDD, all trained on MVTec AD-noise-0.1. The samples, drawn from the test set under *No Overlap* setting, highlight that our method, especially compared to the baseline RD, effectively localizes anomalies.

### F.2 Qualitative Results on VisA

Similar to Fig. 4, we calculate anomaly scores for samples in the training and test set of VisA-noise-0.05, resulting in the anomaly score distribution histograms shown in Fig. A7. These histograms demonstrate that, compared to the baseline RD [9], our proposed method generates anomaly scores that more effectively distinguish between anomalous and normal samples.

Additionally, we compared the anomaly maps of our CDD with SoftPatch [17], InReaCh [24], and the baseline RD [9] on VisA-noise-0.05 in Fig. A8, showing that our method consistently achieves effective anomaly localization across various scenarios.

### F.3 Qualitative Results on Hard Samples

In FUAD tasks, it is challenging for a trained model to localize anomalies for the anomalous samples in the training set, which led to the introduction of the *Overlap* setting. Therefore, we conduct visualizations by generating anomaly maps for the anomalous samples selected from the training set. Figure A9 and Figure A10 compare anomaly maps from SoftPatch [17], the baseline RD [9], and our CDD on MVTec AD-noise-0.1 and VisA-noise-0.05, respectively. The results show that our method, through cross-domain training, effectively mitigates the noise interference from training set and achieves accurate anomaly localization even on training samples.

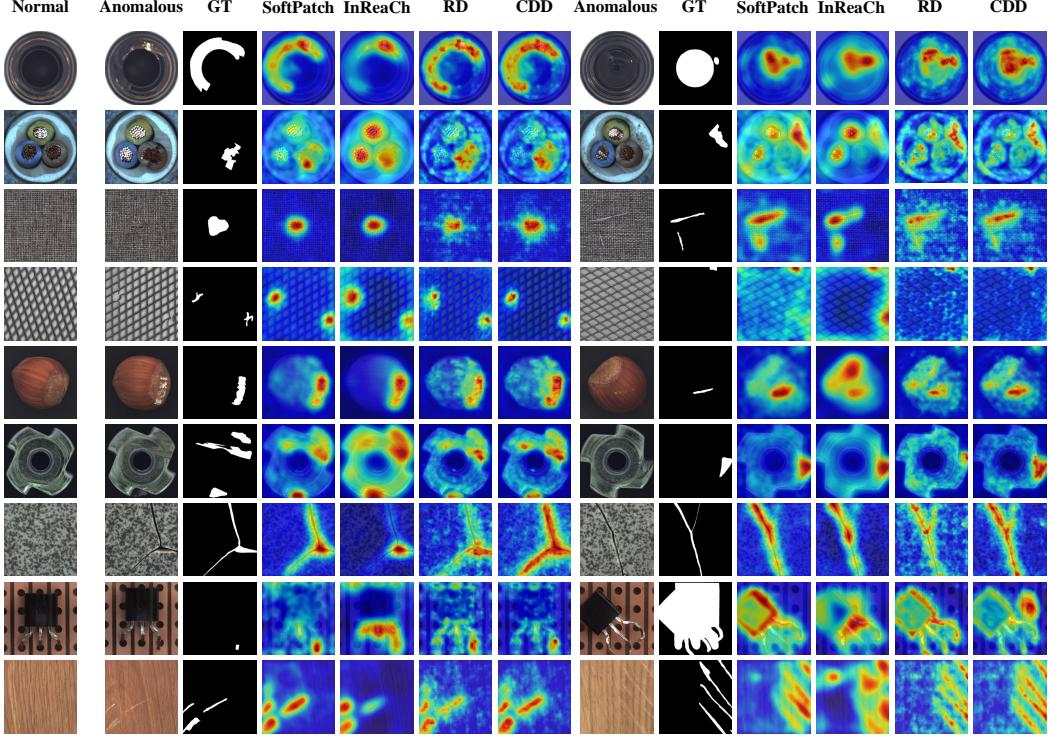


Figure A6: Visualization comparison of anomaly maps generated by FUAD methods SoftPatch [17], InReaCh [24], the baseline RD [9], and our proposed CDD trained on MVTec AD-noise-0.1. All the samples are obtained from the test set under *No Overlap* setting.

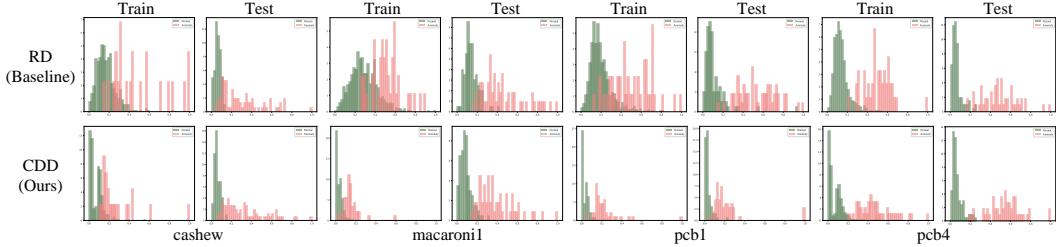


Figure A7: Comparison of histograms of anomaly scores obtained by RD and our CDD. on VisA-noise-0.05.

## G More Discussions

In Sec. 6, we discuss the current limitations of CDD, noting that it has not been extended to more paradigms and remains confined within the RD framework.

In addition, our method currently employs random domain division, which introduces strong randomness and may result in clustering the same anomaly type into a single domain, potentially violating *Assumption 1*. Although *Assumption 2* still allows us to achieve results surpassing the baseline, this randomness somewhat limits the method’s effectiveness.

In future work, we plan to extend this Cross-Domain idea for FUAD to other anomaly detection paradigms, such as feature reconstruction, and apply pre-processing techniques like data clustering before domain construction to ensure the reliability of the process.

## References

- [1] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon. Gandomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth,*

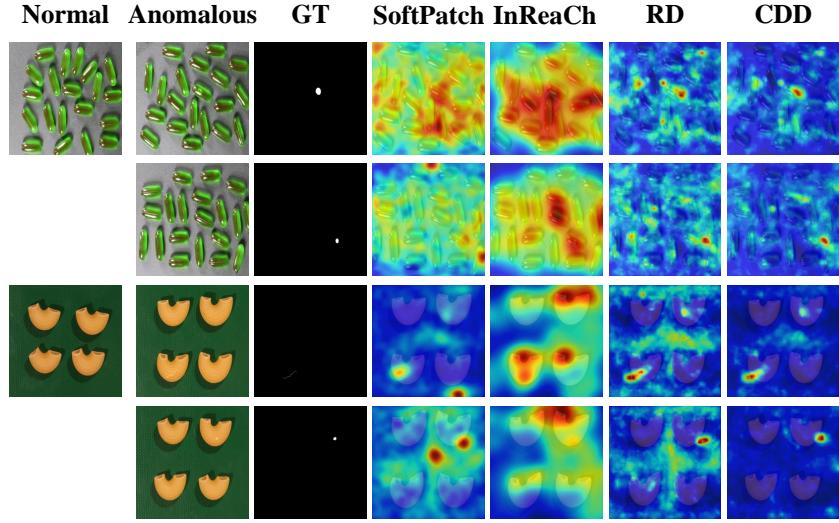


Figure A8: Visualization comparison of anomaly maps generated by FUAD methods SoftPatch [17], InReaCh [24], the baseline RD [9], and our proposed CDD trained on VisA-noise-0.05. All the samples are obtained from the test set under *No Overlap* setting.

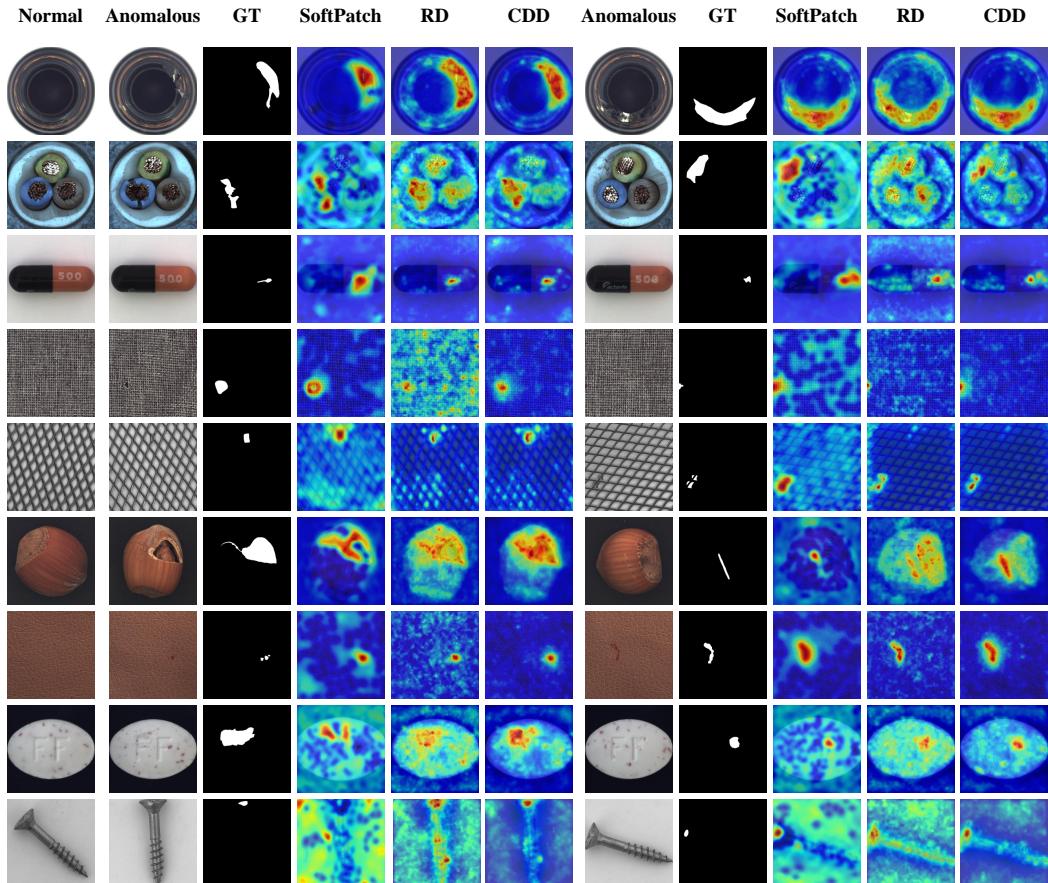


Figure A9: Visualization comparison of anomaly maps generated by FUAD methods SoftPatch [17], the baseline RD [9], and our proposed CDD trained on MVTec AD-noise-0.1. All the anomalous samples are obtained from the train set.

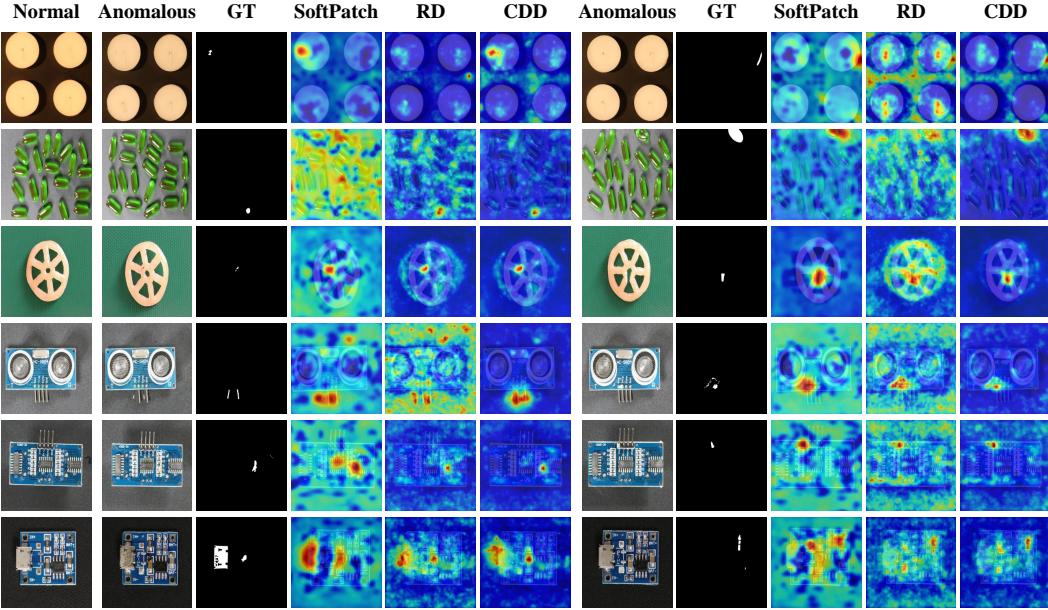


Figure A10: Visualization comparison of anomaly maps generated by FUAD methods SoftPatch [17], the baseline RD [9], and our proposed CDD trained on VisA-noise-0.05. All the anomalous samples are obtained from the train set.

- Australia, December 2–6, 2018, Revised Selected Papers, Part III 14, pages 622–637. Springer, 2019.* 1, 3
- [2] J. Bae, J.-H. Lee, and S. Kim. Pni: industrial anomaly detection using position and neighborhood information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6373–6383, 2023. 1, 3
  - [3] 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*, pages 128–138, 2024. 3
  - [4] 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. 1, 3
  - [5] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger. Mvtac ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 1, 2, 13
  - [6] 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*, pages 4183–4192, 2020. 1, 3
  - [7] 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*, 130(4):947–969, 2022. 1
  - [8] 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*, pages 475–489. Springer, 2021. 3
  - [9] 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*, pages 9737–9746, 2022. 2, 3, 7, 8, 19, 20, 21, 22, 23, 24
  - [10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

- [11] Z. Gu, L. Liu, X. Chen, R. Yi, J. Zhang, Y. Wang, C. Wang, A. Shu, G. Jiang, and L. Ma. Remembering normality: Memory-guided knowledge distillation for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16401–16409, 2023. 3
- [12] D. Gudovskiy, S. Ishizaka, and K. Kozuka. 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*, pages 98–107, 2022. 3
- [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*, pages 6447–6458, 2023. 3
- [14] J. Guo, L. Jia, W. Zhang, H. Li, et al. Recontrast: Domain-specific anomaly detection via contrastive reconstruction. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 19, 20
- [15] J. Hyun, S. Kim, G. Jeon, S. H. Kim, K. Bae, and B. J. Kang. Reconpatch: Contrastive patch representation learning for industrial anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2052–2061, 2024. 3
- [16] J. Im, Y. Son, and J. H. Hong. Fun-ad: Fully unsupervised learning for anomaly detection with noisy training data. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9447–9456. IEEE, 2025. 2, 3, 7, 8, 19, 20, 21
- [17] X. Jiang, J. Liu, J. Wang, Q. Nie, K. Wu, Y. Liu, C. Wang, and F. Zheng. Softpatch: Unsupervised anomaly detection with noisy data. *Advances in Neural Information Processing Systems*, 35:15433–15445, 2022. 2, 3, 7, 8, 19, 20, 21, 22, 23, 24
- [18] Y. Jiang, Y. Cao, and W. Shen. A masked reverse knowledge distillation method incorporating global and local information for image anomaly detection. *Knowledge-Based Systems*, 280:110982, 2023. 3
- [19] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9664–9674, 2021. 1, 3
- [20] H. Li, Z. Chen, Y. Xu, and J. Hu. Hyperbolic anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17511–17520, 2024. 3
- [21] J. Lin and Y. Yan. A comprehensive augmentation framework for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8742–8749, 2024. 3
- [22] X. Liu, J. Wang, B. Leng, and S. Zhang. Dual-modeling decouple distillation for unsupervised anomaly detection. In *ACM Multimedia 2024*, 2024. URL <https://openreview.net/forum?id=TOMVFf5L6Q>. 3
- [23] X. Liu, J. Wang, B. Leng, and S. Zhang. Unlocking the potential of reverse distillation for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5640–5648, 2025. 7, 8, 19, 20
- [24] D. McIntosh and A. B. Albu. Inter-realization channels: Unsupervised anomaly detection beyond one-class classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6285–6295, 2023. 2, 3, 7, 8, 19, 20, 21, 22, 23
- [25] P. Perera, R. Nallapati, and B. Xiang. Ogan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2898–2906, 2019. 3
- [26] 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*, pages 14318–14328, 2022. 1, 3
- [27] 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*, pages 2592–2602, 2023. 2, 3
- [28] 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*, pages 14902–14912, 2021. 2, 3

- [29] 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*, 54:30–44, 2019. [3](#)
- [30] 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*, pages 24511–24520, 2023. [3](#)
- [31] C. Wang, W. Zhu, B.-B. Gao, Z. Gan, J. Zhang, Z. Gu, S. Qian, M. Chen, and L. Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22883–22892, 2024. [3](#)
- [32] G. Wang, S. Han, E. Ding, and D. Huang. Student-teacher feature pyramid matching for anomaly detection. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 306. BMVA Press, 2021. [2](#), [3](#)
- [33] Z. You, L. Cui, Y. Shen, K. Yang, X. Lu, Y. Zheng, and X. Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022. [3](#)
- [34] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. [5](#)
- [35] 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*, pages 8330–8339, 2021. [1](#), [3](#)
- [36] 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*, pages 149–158, 2024. [3](#)
- [37] X. Zhang, N. Li, J. Li, T. Dai, Y. Jiang, and S.-T. Xia. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6782–6791, 2023. [3](#)
- [38] X. Zhang, M. Xu, and X. Zhou. 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*, pages 16699–16708, 2024. [3](#)
- [39] Y. Zhou, X. Xu, J. Song, F. Shen, and H. T. Shen. Msflow: Multiscale flow-based framework for unsupervised anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. [3](#)
- [40] 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*, pages 392–408. Springer, 2022. [17](#)