Prototype-Aware Contrastive Knowledge Distillation for Few-Shot Anomaly Detection

Zhihao Gu[1]
ellery-holmes@sjtu.edu.cn
Taihai Yang[2]
thyang@stu.ecnu.edu.cn
Lizhuang Ma[†1,2]
ma-lz@cs.sjtu.edu.cn

[1] School of EIEE
Shanghai Jiao Tong University
Shanghai, China

[2] East China Normal University
Shanghai, China

Abstract

Knowledge distillation (KD) is widely adopted in anomaly detection but how to extend it to the few-shot setting, where a few normal samples are provided for detecting anomalies in unseen categories, has not been explored yet. To remedy this problem, we propose a novel ***P**rototype-**A**ware **C**ontrastive **K**nowledge **D**istillation* (**PACKD**) framework. Specifically, we first design a prototype extraction and integration module (PEIM) to improve the generalization of the KD model by integrating prior information of a given category from the teacher network into the student network. The PEIM is trained to generate prototypes from few-shot normal samples to give priors and further uses them to guide the student to restore distillation targets. Subsequently, we adopt a novel contrastive distillation strategy to robustly distill both normal sample representations and inter-sample relations in the training phase. The negative and positive pairs are obtained from the feature correlations of the teacher and student. Comprehensive studies demonstrate that the proposed method outperforms the comparable few-shot methods on three benchmarks, even in more challenging cross-dataset scenarios.

1 Introduction

Anomaly detection (AD) receives quite some attention in recent years due to its wide range of applications, like defect detection [2], video surveillance [11] and medical diagnosis [43]. Since it is difficult to collect an exhaustive set of anomalous samples, recent efforts [8, 21, 28, 29, 37, 40] usually formulate it as an unsupervised learning problem (vanilla AD), where only normal data is available, and has developed into several categories: reconstruction [14, 32, 35, 41], knowledge distillation (KD) [2, 3, 9, 30], embedding [7, 8, 19, 27] and generation [20, 21, 39, 42]. They model normal distribution from the training data and samples that deviated from the distribution are considered as anomalies. Nevertheless, these approaches need abundant data to train a category-dependent model for each class, which is inefficient in real-world scenarios like defect detection. To reduce the demand for training samples, a couple of studies tend to explore few-shot anomaly detection (FSAD), which detects anomalies in a target category with a handful of normal samples.

[†]Corresponding author. Work done when Zhihao Gu is an intern at CATL.

Most existing FSAD methods focus on extending approaches in vanilla AD to few-shot scenarios. For instance, PatchCore [27] and DifferNet [28] are directly evaluated in the few-shot setting. RegAD [15] introduces a proxy task to C2F [44] and trains a category-agnostic model for AD. FAAD [37] proposes an adaptive sparse coding layer for EBAD [10] to avoid retraining for new categories. However, we observe that knowledge distillation, one of the mainstream methods in vanilla AD, has not been explored in the few-shot setting. To remedy this problem, in this paper, we integrate two novel components into reverse knowledge distillation (RD) [9] for few-shot anomaly detection.

The fundamental challenge lies in that the scarcity of the target category data makes fine-tuning challenging. To mitigate this issue, we formulate the FSAD as a meta-learning problem and propose a novel ***P**rototype-**A**ware **C**ontrastive **K**nowledge **D**istillation* (**PACKD**) framework that improves the generalization by exploring discriminative information from few-shot normal samples. Specifically, we first devise a prototype extraction and integration module (PEIM) to extract prior information of a given category and guide the reconstruction of distillation targets. The PEIM is trained to generate prototypes from the teacher representations of support images and integrates these priors into the student representations of the query image for decoding. The student thus generalizes to unseen categories. Then we adopt a novel contrastive distillation strategy (CDS) to constrain the reconstruction results between categories during training. Given the teacher and student representation of a query sample, the teacher representations of its support sample are selected as anchors. The feature correspondences of the anchor-student are encouraged to be consistent with that of anchor-teacher, which further guarantees the robustness of knowledge distillation. To the best of our knowledge, PACKD is the first KD-based method for few-shot anomaly detection.

Contributions. (1) We present a novel *Prototype-Aware Contrastive Knowledge Distillation* paradigm that explores KD in FSAD. (2) We devise *a prototype extraction and integration module* to improve the generalization of the KD model by integrating prior information. (3) We adopt a novel *contrastive distillation strategy* to improve the robustness of KD. (4) The proposed method outperforms the comparable methods in different few-shot scenarios.

2 Related Works

Few-Shot Anomaly Detection. Anomaly detection has achieved prominent progress in the past few decades and can be categorized into several groups: reconstruction-based methods [9, 30, 35, 41], feature embedding [7, 8, 19, 27] and generation-based approaches [20, 21, 39, 42]. They have to collect hundreds of data to train a dedicated model for each given category, which is time-consuming and inefficient. To improve efficiency, the few-shot setting begins to achieve attention recently, where only limited samples are provided to detect anomalies. TDG [31] trains the model by distinguishing which transformation is applied to image patches and patch-based votes of correct transformation give the anomaly score. RegAD [15] aggregates normal data from different categories to train a model and anomalies in a new category are identified by comparing features between test and support images. FAAD [37] models abnormal distribution and designs a sparse coding layer for model adaptation. Energy-based models are then leveraged to detect anomalies. In this paper, following the setting in [15, 37], we explore knowledge distillation for few-shot anomaly detection, which has not been explored yet.

Knowledge Distillation [13] is originally designed to transfer knowledge from a heavy network to a lightweight one for model compression and has been extensively explored in unsupervised anomaly detection [2, 9, 29, 30, 34]. The discrepancies in features between the teacher and student (S-T) are used for AD. However, they ignore the inter-category relation and KD models are hard to be fine-tuned with limited data. Instead, we formulate the FSAD as a meta-learning problem and design the PEIM to improve the model generalization and a novel contrastive distillation strategy to explore more robust distillation.

Memory Networks. Recently, memory-augmented neural networks have been introduced in various computer vision fields [6, 11, 14, 16, 36]. For example, MID [6] designs a memory network to capture rain streak information in time-lapse data. CDFSS [36] proposes a memory bank to store the style from source domain instances for enhancing target samples. In the context of anomaly detection, the memory mechanism [11, 14, 24, 26] constructs memory banks to store normal patterns during training for suppressing the generalization of Auto-Encoders [18]. In the few-shot setting, it is hard to build a memory bank from limited samples. We instead train a lightweight network to adaptively generate representative features (prototypes) from few-shot normal samples of novel categories, which are further used to guide the student to restore targets of distillation.

Contrastive Learning [4, 12, 25, 38] aims to learn visual representations via attracting similar instances while repelling dissimilar ones. Some recent works [23, 33, 45] introduce it to anomaly detection. For example, CRADL [23] learns more semantic-rich representations by it to fix the over-fixation of low-level features. SPD [45] proposes the SmoothBlend to produce negatives and treats globally augmented images as positives for conducting contrastive learning. Differently, we adopt a novel contrastive distillation strategy for KD to explore the knowledge of intra and inter-sample relations during the training stage.

3 Problem Formulation

We follow previous works [15, 37] to formalize the FSAD as a meta-learning problem, where the model is trained on several categories while tested on unseen/novel categories.

Assume the training set consists of N categories, *i.e.*, $\bigcup_{c=1}^{N} \mathcal{T}_c = \{(X_q, \{X_s^n\}_{n=1}^{k})_j\}_{j=1}^{|\mathcal{T}_c|}$, where $X_q \in R^{H \times W}$ refers to the query image and $\{X_s^n\}_{n=1}^{k}$ is its corresponding k normal images (support images). The PEIM is trained to extract prior information from $\{X_s^n\}_{n=1}^{k}$ and integrate them into X_q. Then the CDS is applied to representations of X_q and loss is calculated. Each test sample of novel categories also owns k normal samples in inference. The PEIM adopts the same process to them as mentioned above. Finally, anomaly detection is conducted based on the test image following the vanilla AD [9]. Next, we will focus on the 1-shot setting and how to extend it to the $k(k > 1)$-shot setting is described in Sec. 4.4.

4 Prototype-Aware Contrastive Knowledge Distillation

In this section, we present the *Prototype-Aware Contrastive Knowledge Distillation* (**PACKD**) that explores the RD [9] in the few-shot setting, as shown in Fig. 1. The main idea is to use information in the support image to guide anomaly detection on the query sample. To this end, we first devise the prototype extraction and integration module to improve the model generalization by integrating prior information of the support image from the teacher network into the student network. Then a new contrastive distillation strategy is adopted to explore the knowledge of intra and inter-sample relations for more robust knowledge distillation.

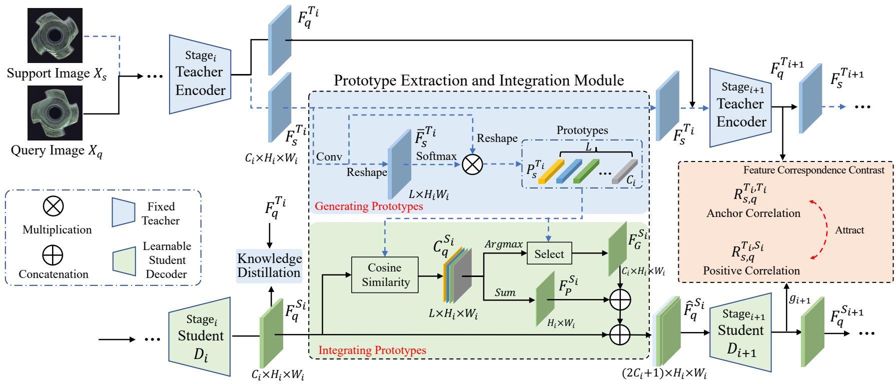

Figure 1: Overview of the proposed PACKD framework. We first use the prototype extraction and integration module to extract prior information from the teacher representation of the support image. These priors are then integrated into the student representation of the query to guide the reconstruction of the distillation target. Moreover, we adopt a contrastive loss to constrain the retrieved representation from the feature correlation perspective.

4.1 Preliminaries: Reverse Knowledge Distillation for FSAD

Reverse knowledge distillation (RD) [9] is a recently proposed method for vanilla AD and we choose it as our basic paradigm due to its efficiency. It owns a pre-trained encoder (the teacher network) and a trainable decoder (the student network), which is built on the one-class embedding of the teacher. In the few-shot setting, given a normal sample for reference, the RD is expected to detect anomalies on the test sample of the same category.

In the training phase, given a support sample $X_s \in R^{C \times H \times W}$ and the query sample $X_q \in R^{C \times H \times W}$ of the target category, the teacher extracts features $\{F_s^{T_i}\}_{i=1}^3 \in R^{C_i \times H_i \times W_i}$ based on X_s from the first three stages and the student gives corresponding representations $\{F_s^{S_i}\}_{i=1}^3$, where C_i, H_i and W_i are the channel, height, and width at i^{th} stage, respectively. Then a knowledge distillation loss is used to enforce the feature consistency between them:

$$\mathcal{L}_{\text{KD}}(F_s^{T_i}, F_s^{S_i}) = 1 - \frac{\text{flat}(F_s^{T_i})}{\|\text{flat}(F_s^{T_i})\|_2} \cdot \frac{\text{flat}(F_s^{S_i})^\top}{\|\text{flat}(F_s^{S_i})\|_2}, \tag{1}$$

where $\text{flat}(\cdot) : R^{C_i \times H_i \times W_i} \rightarrow R^{C_i H_i W_i}$ is the flatten function and $\|\cdot\|$ means the l_2 norm. In inference, the pixel-wise similarity between $\{F_q^{T_i}, F_q^{S_i}\}_{i=1}^3$ is computed for anomaly detection.

4.2 Prototype Extraction and Integration Module (PEIM)

The scarcity of support images of the target category makes training RD challenging. Thus, the FSAD is formulated as a meta-learning problem. Then how to use the support image becomes vital. Note that the student aims to restore the teacher's representations. So introducing information about novel categories to the student is beneficial. To this end, we design a prototype extraction and integration module to extract priors from teacher representations of the support sample and later integrate them into student representations of the query.

Generating prototypes. A direct way is to store features of the support image. However, the size of the features makes it inefficient. We instead train a lightweight network to adaptively generate a fixed number of prototypes from the teacher representation of the support image.

Concretely, given the feature of the support image $F_s^{T_i} = \{(F_s^{T_i})^m \in R^{C_i}\}_{m=1}^{H_iW_i}$ at i^{th} stage, we first apply a convolution ($R^{C_i \times H_i \times W_i} \rightarrow R^{L \times H_i \times W_i}$) of kernel size 3×3 and a reshape operation on it to produce feature $\bar{F}_s^{T_i} = \{(\bar{F}_s^{T_i})^l \in R^{H_i \times W_i}\}_{l=1}^{L}$. Then the softmax is conducted on $(\bar{F}_s^{T_i})^l$ to give L attention map for aggregating the spatial dimension of $F_s^{T_i}$ and generating L prototypes $P_s^{T_i} = \{(P_s^{T_i})^l \in R^{C_i}\}_{l=1}^{L}$. The whole process can be formulated as follows:

$$(P_s^{T_i})^l = \sum_{m=1}^{H_iW_i} \frac{(\bar{F}_s^{T_i})^{l,m}}{\sum_{h,w}(\bar{F}_s^{T_i})^{l,m}} \cdot (F_s^{T_i})^m, \tag{2}$$

where (h,w) indicates the spatial index. We use the orthogonal loss L_{orth} [36] to make sure that each prototype is as independent as possible from others. The generated prototypes contain class-specific information from normal samples and we consider them as the priors.

Integrating prototypes. Since the number of support images is limited, we provide each location in the query feature with the most similar prototype and their similarity to L prototypes, which is different from previous works [11, 24, 26] using prototypes for retrieval.

Formally, we first measure the cosine similarity $C_q^{S_i} = \{(C_q^{S_i})^l \in R^{H_i \times W_i}\}_{l=1}^{L}$ between each prototype $(P_s^{T_i})^l$ and each location (h,w) on the query feature $F_q^{S_i} \in R^{C_i \times H_i \times W_i}$ as follows:

$$(C_q^{S_i})_{h,w}^l := \mathrm{Sim}((P_s^{T_i})^l, (F_q^{S_i})_{h,w}) = \sum_c \frac{(P_s^{T_i})^l \cdot (F_q^{S_i})_{h,w}}{\|(P_s^{T_i})^l\| \cdot \|F_q^{S_i})_{h,w}\|}. \tag{3}$$

Then, for each position, a prototype $(P_s^{T_i})^{l_{h,w}}$ with the largest similarity are selected to form the guide feature $F_G^{S_i} \in R^{C_i \times H_i \times W_i}$, where $l_{h,w} = \mathrm{argmax}_l (C_q^{S_i})_{h,w}^l$. We also add up similarity information in $C_q^{S_i}$ across all prototypes to get the probability map $F_P^{S_i} \in R^{H_i \times W_i}$. Finally, the original query feature, the guide feature, and the probability map are concatenated along channel dimension to provide guiding information for the student D_{i+1}, resulting in $F_q^{S_{i+1}}$:

$$F_q^{S_{i+1}} = D_{i+1}(F_q^{S_i} \oplus F_G^{S_i} \oplus F_P^{S_i}), \tag{4}$$

where $\oplus$ is the concatenation. $F_q^{S_{i+1}}$ is enforced to be consistent with $F_q^{T_{i+1}}$ by Eq. (1).

4.3 Contrastive Distillation Strategy (CDS)

Recent progress [22] in contrastive learning has yielded methods that empower the representation of few-shot models. Inspired by this, we conduct the contrastive loss to explore rich information from intra- and inter-instance correlations for KD. The teacher and student representations of the support image and the query are exploited to form contrastive pairs.

We define the correspondence $R_{a,b}^{T_i,S_i} \in R^{H_iW_i \times H_iW_i}$ between two features $F_a^{T_i}, F_b^{S_i} \in R^{C_i \times H_i \times W_i}$:

$$R_{h_aw_ah_bw_b}^{T_i,S_i}(F_a^{T_i}, F_b^{S_i}) = \sum_c \frac{(F_a^{T_i})_{h_a,w_a} \cdot (F_b^{S_i})_{h_b,w_b}}{\|(F_a^{T_i})_{h_a,w_a}\| \cdot \|(F_b^{S_i})_{h_b,w_b}\|}, \tag{5}$$

where each entry stands for the cosine similarity between feature at spatial position (h_a,w_a) of $F_a^{T_i}$ and position (h_b,w_b) of $F_b^{S_i}$. In the context of knowledge distillation, the correspondence between teacher and student features provides semantic relations of different regions from normal samples and we adopt it to construct the contrastive objective. More specifically, given a query's teacher and student representations $F_q^{T_i}$ and $g_i(F_q^{S_i})$ at i^{th} stage, where $g_i(\cdot)$ is the projection head [5] for transformation, we select the correspondence between teacher representation $F_s^{T_i}$ of the support image and $F_q^{T_i}$ as an anchor correlation, denoted as $R_{s,q}^{T_i,T_i}$. $R_{s,q}^{T_i,S_i}$ and $R_{s,q^-}^{T_i,S_i}$ are treated as positive-negative pairs, where q^- belongs to a different category. The contrastive distillation loss with the temperature coefficient τ is formulated as:

$$\mathcal{L}_{\text{NCE}}(F_s^{T_i}, F_q^{T_i}, F_q^{S_i}) = -\, R_{s,q}^{T_i,T_i} \cdot R_{s,q}^{T_i,S_i}/\tau + \log(\sum_{q^-} e^{\,R_{s,q}^{T_i,T_i} \cdot R_{s,q^-}^{T_i,S_i}/\tau}). \tag{6}$$

The CDS attracts intra-sample correlations while repelling inter-sample ones. And this formulation brings several merits. 1) Since the sources for calculating the anchor are fixed, the student can be effectively optimized. 2) As one query sample owns k support images, the student is supervised by multiple anchors, which further guarantees the robustness of KD.

4.4 Extension to the k-shot Setting and Anomaly Detection

Extension to the k-shot setting. In PEIM, we generate L prototypes for each support sample and put these kL prototypes together for similarity calculation in Eq. (3) and further selection. Besides, for the CDS, Eq. (6) is computed on all support samples and we take their average.

Finally, the objective contains the distillation loss, contrastive loss and orthogonal loss:

$$\mathcal{L} = \sum_{i=1}^{3} [\mathcal{L}_{\text{KD}}(F_q^{T_i}, F_q^{S_i}) + \frac{\lambda_1}{k} \cdot \sum_{n=1}^{k} \mathcal{L}_{\text{NCE}}(F_{s_n}^{T_i}, F_q^{T_i}, F_q^{S_i}) + \lambda_2 \cdot L_{orth}(P_s^{T_i})], \tag{7}$$

where λ_1 and λ_2 are a balancing hyper-parameter and set 0.5 by default.

5 Experiments

5.1 Experimental Setup

Datasets. Our experiments are based on three large-scale benchmarks, *i.e.,* MVTec AD [1], VisA [45], MPDD [17]. The MVTec AD contains more than 5000 images of 15 classes and the Visa is composed of 10,821 images for 12 categories. Besides, MPDD consists of 6 classes of about 1300 images. Images in these benchmarks own full pixel-level annotations.
Evaluation metrics. We evaluate our method by the Area Under the Receiver Operator Curve (AUROC), which is a common metric adopted in AD [9, 15]. The image-level and pixel-level AUROC (%) are computed for anomaly detection and localization, respectively.
Baselines. We compare the proposed PACKD with several SOTA few-shot and vanilla AD methods. For the former, DifferNet [28], TDG [31], PatchCore [27] and RegAD [15] are selected, which are trained by their default settings. For the latter, we consider SPADE [7], STPM [34] and RD [9] and train them with the provided normal samples of novel categories.
Implementation details. All images are resized into 256×256 and Adam is used as the optimizer with a learning rate of 0.0005 during training. The model is trained for 100 epochs with a batch size of 32. L is set to 20 for all stages and τ is 0.05. We employ the default settings in [9] to implement RD, *i.e.,* an ImageNet pre-trained WideResNet50 as the teacher and a corresponding reversed structure as the student network. Support samples are augmented by rotation as in [15] and all experiments are conducted on a Nvidia Tesla V100 GPU.

Method	MVTec AD [1]			VisA [45]			MPDD [17]		
	k=2	k=4	k=8	k=2	k=4	k=8	k=2	k=4	k=8
SPADE [7]	70.7	71.6	75.3	80.7	81.7	-	58.2	58.3	58.5
STPM [34]	74.2	74.8	77.6	72.5	73.1	74.6	62.4	62.6	63.1
RD [9]	75.5	76.9	78.5	75.9	76.9	77.2	61.8	62.1	62.4
TDG [31]	73.2	74.4	76.7	71.7	72.9	73.4	60.3	63.5	68.2
DifferNet [28]	80.6	81.3	83.2	74.3	76.2	80.6	60.2	63.3	68.5
PatchCore [27]	87.8	89.5	94.3	81.6	85.3	-	59.5	59.8	60.0
RegAD [15]	85.7	88.2	91.2	75.3	79.5	81.9	63.4	68.3	**71.9**
Ours	**90.2**	**91.6**	**95.3**	**83.4**	**86.3**	**87.5**	**66.6**	**69.8**	70.5

(a) K-shot anomaly detection performance. Image-level AUROC (%) is reported.

Method	MVTec AD [1]			VisA [45]			MPDD [17]		
	k=2	k=4	k=8	k=2	k=4	k=8	k=2	k=4	k=8
SPADE [7]	79.8	80.2	80.5	82.2	82.6	-	75.4	75.9	76.2
STPM [34]	59.8	60.8	61.6	71.4	72.3	73.1	75.8	76.2	76.6
RD [9]	79.3	81.4	83.9	79.3	81.4	81.9	74.5	75.5	75.7
PatchCore [27]	91.0	91.3	91.6	94.7	95.7	96.3	78.2	78.7	79.0
RegAD [15]	94.6	95.8	96.8	95.6	96.0	97.0	93.2	93.9	95.1
Ours	**95.0**	**96.2**	**97.3**	**96.3**	**97.2**	**97.9**	**94.4**	**94.8**	**95.3**

(b) K-shot anomaly localization performance. Pixel-level AUROC (%) is reported.

Table 1: Few-shot anomaly detection results on MVTec AD, VisA, and MPDD datasets. Results are the average score over all categories and listed as the average AUC of 10 runs.

5.2 Main Results

Few-shot anomaly detection and localization. Tab. 1 (a) and (b) respectively demonstrate the comprehensive comparisons of k-shot anomaly detection and localization on MVTec AD [1], VisA [45] and MPDD [17] benchmarks. Several representative works in FSAD and vanilla AD are studied. It is found that a larger k leads to better results since more information about the category is provided. Besides, methods in the vanilla AD, *i.e.*, the RD, STPM, and SPADE under-perform their competitors in the few-shot setting. This derives from the fact that compared to FSAD approaches with fewer trainable parameters, the scarcity of category data makes fine-tuning these models with massive parameters challenging. On the contrary, the proposed PACKD is trained to explore discriminative information from few-shot novel categories for better generalization and thus achieves better results.

Cross-dataset FSAD. In real-world industrial scenarios, there exist domain shifts, derived from varying poses and imaging conditions, between novel categories and the seen ones. To model this setting, we pre-train the proposed method on one dataset and test it on another one. Tab. 2 demonstrates the results on the MVTec AD and MPDD datasets. Compared to the intra-dataset setting in Tab. 1, domain shifts make AD more difficult and thus degrade the overall performance. Moreover, results from pre-training on MPDD outperform those from pre-training on MVTec AD by about 20% because the MPDD dataset is more challenging for its various spatial orientations, light intensities, and non-homogeneous backgrounds. And the learned ability on it can be adapted to detecting anomalies in easier situations.

5.3 Ablation study

We conduct ablation studies to evaluate the proposed method and the RD [9] is our baseline.

Method	MVTec AD→MPDD			MPDD→MVTec AD		
	k=2	k=4	k=8	k=2	k=4	k=8
TDG [31]	54.3	59.1	60.9	66.2	67.2	68.6
DifferNet [28]	57.6	62.2	66.4	76.0	78.9	82.8
RegAD [15]	60.1	63.4	67.6	82.4	85.1	88.7
Ours	**62.1**	**65.8**	**68.4**	**83.8**	**86.9**	**90.5**

Table 2: Cross-dataset few-shot anomaly detection results on MVTec AD and MPDD benchmarks. Test samples come from a different dataset and image-level AUROC (%) is reported.

PEIM	CDS	k=2	k=4
		75.5	76.9
√		86.1	89.0
	√	82.7	84.3
√	√	**90.2**	**91.6**

(a) Key components on MVTec AD.

k	MVTec AD	VisA
2	90.2	83.4
8	95.3	87.5
32	95.9	89.7
64	**96.8**	**93.6**

(b) Number of shots.

N	$k=2$	$k=4$
0	75.5	76.9
5	81.5	84.9
10	85.7	88.2
14	**90.2**	**91.6**

(c) Training categories.

L	$k=2$
10	89.6
20	90.2
50	90.5
100	**90.9**

(d) Prototypes.

Backbone	MVTec AD	VisA	MPDD
RegAD [15]	85.7	75.3	63.4
ResNet18	86.7	77.3	63.8
ResNet34	87.6	80.2	64.5
ResNet50	88.9	82.1	65.4
WideResNet50	**90.2**	**83.4**	**66.6**

(e) Different teacher backbones.

Operation	MVTec AD	VisA	MPDD
Generation [11]	82.7	77.9	62.4
Integration [11]	88.4	82.2	65.2
Ours	**90.2**	**83.4**	**66.6**
w/o rotation [15]	87.2	81.1	63.9

(f) Prototype generation and integration.

Table 3: Ablation study on different benchmarks. Image-level AUROC (%) is reported.

Study on key components. We study the impacts of the prototype extraction and integration module (PEIM) and the contrastive distillation strategy (CDS). Tab. 3 (a) reports the results. The baseline (first row) having no access to the test categories owns inferior performance. Introducing priors via the PEIM significantly improves the baseline by about 10.6%. Compared to PEIM, the CDS gives less improvement (7.2% *v.s.* 10.6% on $k=2$ and 7.4% *v.s.* 12.1% on $k=4$), which means learning the ability to extract and integrate priors is more important. Of course, combining them all achieves the best results.

Study on the number of shots. The shot k controls the amount of the prior information about categories and we investigate its effects in Tab. 3 (b). It is observed that a larger k contains more priors and thus better results are obtained. However, compared to the VisA dataset, the performance gains for the MVTec AD dataset are limited as k increases, *e.g.*, 0.9% $\uparrow$ versus 3.9% $\uparrow$ from $k=32$ to $k=64$. We guess samples in the same category, *e.g.*, the screw, are similar and thus provide redundant information, limiting the improvement.

Study on category number. We train the model on different categories to obtain the ability of extracting prior information from few-shot normal samples and integrating them to the test sample. Impacts of the category number are explored in Tab. 3 (c). First of all, adopting the pre-training consistently improves anomaly detection. Besides, more categories give more cases to model the extraction and integration process, which produces better results.

Study on prototype number. The number of prototypes decides how many priors are extracted from support samples. Tab. 3 (d) shows its effects. As can be found that generating more prototypes brings better results. However, the performance gains are not proportional to the increased amount of L (0.6% $\uparrow$ from $L=10 \rightarrow L=20$, 0.3% $\uparrow$ from $L=20 \rightarrow L=30$ and so on). To balance the performance and time consumption, L is set to 20 by default.

Study on prototype generation and integration. In the paper, we train a lightweight network to generate and integrate prototypes. Methods in [11, 14] can also be used for this purpose, which is explored in the first three rows of Tab. 3 (f). Since they set memory vectors as parameters of the network and optimize them by back-propagation, rich information

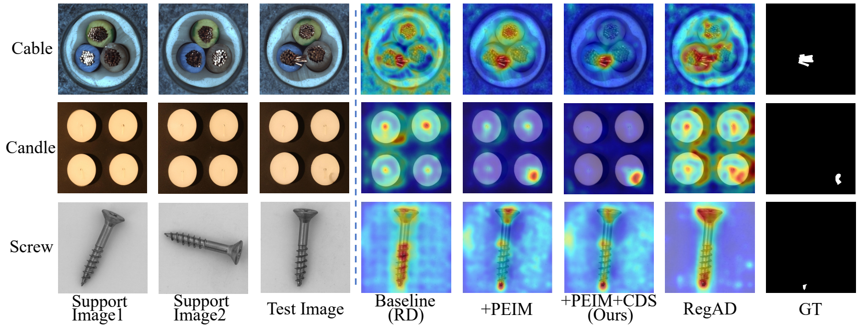

Figure 2: Visualization on impacts of PEIM and CDS for anomaly detection in the 2-shot setting. The PEIM pays attention to anomalous regions while CDS further suppresses responses to anomaly-free areas. Our method localizes anomalies more accurately than RegAD [15].

in the support sample is not explored and inferior results are obtained. Instead, our method is trained to generate and integrate priors from the support sample to provide the network with guidance for detecting anomalies on the query sample, leading to better performance.
Study on backbones. Tab. 3 (e) gives the ablation on backbones. The WideResNet50, which is deeper and wider, has a stronger representative capacity and thus facilitates the detection of anomalies. Besides, compared to the work RegAD [15], building the proposed PACKD upon smaller neural networks, *e.g.*, ResNet18 and ResNet34, still owns competitive performance.
Study on the support set augmentation. Following previous works [15, 28, 31], we also adopt the rotation transformation to few-shot samples for augmentation. Tab. 3 (f) investigates its impacts. We observe that it produces consistent improvements, *i.e.,* 3.0% $\uparrow$ on MVTec AD, 2.3% $\uparrow$ on VisA, and 2.7% $\uparrow$ on MPDD. Since the augmentation diversifies few-shot samples, the PEIM extracts richer prior information from these augmented data to provide guidance for the student network, thus benefiting anomaly detection.

5.4 Visualization

To Intuitively illustrate how the proposed PEIM and CDS improve the baseline RD [9] for AD, we give some visual comparisons in Fig. 2. RD presents poor generalization since the tested categories are unseen during training. Taking support images as a reference, the PEIM provides vital cues for detecting anomalies in the test image. For example, the wire of the “cable" in support images is straight but it is bent in the test image. Similar cases can be found in the “candle” and “screw”. It is also observed that the CDS helps suppress responses to anomaly-free areas since it ensures the feature consistency between the S-T for them.

6 Conclusion

In this work, we present a novel Prototype-Aware Contrastive Knowledge Distillation framework to explore knowledge distillation in few-shot anomaly detection. A prototype extraction and integration module is first proposed to generate prototypes from the teacher representations of support images and integrate these priors into the student representations of the query image for later decoding, significantly improving the generalization. Then, a novel contrastive distillation strategy is adopted to further improve the robustness of KD.

Acknowledgements

This research is supported in part by National Natural Science Foundation of China (No. 61972157 and 72192821), Shanghai Municipal Science and Technology Major Project (2021 SHZDZX0102), Shanghai Science and Technology Commission (21511101200), Shanghai Sailing Program (22YF1420300 and 23YF1410500), CCF-Tencent Open Research Fund (RAGR20220121) and Young Elite Scientists Sponsorship Program by CAST (2022QNRC001).

References

[1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten 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 4182–4191, 2020.

[3] Tri Cao, Jiawen Zhu, and Guansong Pang. Anomaly detection under distribution shift. *arXiv preprint arXiv:2303.13845*, 2023.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607, 2020.

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[6] Jaehoon Cho, Seungryong Kim, and Kwanghoon Sohn. Memory-guided image deraining using time-lapse data. *IEEE Transactions on Image Processing*, 31:4090–4103, 2022.

[7] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020.

[8] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition.*, volume 12664, pages 475–489, 2020.

[9] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9727–9736, 2022.

[10] Ergin Utku Genc, Nilesh Ahuja, Ibrahima J Ndiour, and Omesh Tickoo. Energy-based anomaly detection and localization. *arXiv preprint arXiv:2105.03270*, 2021.

[11] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[13] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[14] Jinlei Hou, Yingying Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, and Hong Zhou. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8771–8780, 2021.

[15] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *Proceedings of the European Conference on Computer Vision*, pages 303–319, 2022.

[16] Somi Jeong, Youngjung Kim, Eungbean Lee, and Kwanghoon Sohn. Memory-guided unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6558–6567, 2021.

[17] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. 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)*, pages 66–71, 2021.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[19] Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10: 78446–78454, 2022.

[20] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas 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.

[21] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. *arXiv preprint arXiv:2303.15140*, 2023.

[22] Yuning Lu, Liangjian Wen, Jianzhuang Liu, Yajing Liu, and Xinmei Tian. Self-supervision can be a good few-shot learner. In *Proceedings of European Conference on Computer Vision*, pages 740–758, 2022.

[23] Carsten T Lüth, David Zimmerer, Gregor Koehler, Paul F Jaeger, Fabian Isensee, Jens Petersen, and Klaus H Maier-Hein. Cradl: Contrastive representations for unsupervised anomaly detection and localization. *arXiv preprint arXiv:2301.02126*, 2023.

[24] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15425–15434, 2021.

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[26] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14360–14369, 2020.

[27] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter V. Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14298–14308, 2022.

[28] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1906–1915, 2021.

[29] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Asymmetric student-teacher networks for industrial anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2591–2601, 2023.

[30] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid 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.

[31] Shelly Sheynin, Sagie Benaim, and Lior Wolf. A hierarchical transformation-discriminating generative model for few shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8475–8484, 2021.

[32] Xijuan Sun, Di Wu, Arnaud Zinflou, and Benoit Boulet. Anomaly detection with ensemble of encoder and decoder. *arXiv preprint arXiv:2303.06431*, 2023.

[33] Gaoang Wang, Yibing Zhan, Xinchao Wang, Mingli Song, and Klara Nahrstedt. Hierarchical semi-supervised contrastive learning for contamination-resistant anomaly detection. In *Proceedings of the European Conference on Computer Vision*, pages 110–128, 2022.

[34] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for anomaly detection. In *British Machine Vision Conference*, page 306, 2021.

[35] Shenzhi Wang, Liwei Wu, Lei Cui, and Yujun Shen. Glancing at the patch: Anomaly localization with global and local feature comparison. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2021.

[36] Wenjian Wang, Lijuan Duan, Yuxi Wang, Qing En, Junsong Fan, and Zhaoxiang Zhang. Remember the difference: Cross-domain few-shot semantic segmentation via meta-memory transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7065–7074, 2022.

[37] Ze Wang, Yipin Zhou, Rui Wang, Tsung-Yu Lin, Ashish Shah, and Ser Nam Lim. Few-shot fast-adaptive anomaly detection. *Advances in Neural Information Processing Systems*, 35:4957–4970, 2022.

[38] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[39] Xuan Xia, Weijie Lv, Xing He, Chuanqi Liu, and Ning Ding. Fractalad: A simple industrial anomaly segmentation method using fractal anomaly generation and backbone knowledge distillation. *arXiv preprint arXiv:2301.12739*, 2023.

[40] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *arXiv preprint arXiv:2206.03687*, 2022.

[41] Vitjan Zavrtanik, Matej Kristan, and Danijel 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.

[42] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dsr–a dual subspace re-projection network for surface anomaly detection. In *Proceedings of the European Conference on Computer Vision*, pages 539–554, 2022.

[43] Jianpeng Zhang, Yutong Xie, Yi Li, Chunhua Shen, and Yong Xia. Covid-19 screening on chest x-ray images using deep learning based anomaly detection. *arXiv preprint arXiv:2003.12338*, 27:141, 2020.

[44] Ye Zheng, Xiang Wang, Rui Deng, Tianpeng Bao, Rui Zhao, and Liwei Wu. Focus your distribution: Coarse-to-fine non-contrastive learning for anomaly detection and localization. In *2022 IEEE International Conference on Multimedia and Expo*, pages 1–6, 2022.

[45] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 392–408, 2022.