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

Zhihao Gu
ellery-holmes@sjtu.edu.cn
Taihai Yang
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 Prototype-Aware Contrastive Knowledge Distillation (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 [1], video surveillance [21] 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.

© 2023. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.
† 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 Prototype-Aware Contrastive Knowledge Distillation (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 un-
supervised 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., \(\bigcup_{c=1}^{N} T_c = \{(X_q, \{X^n_s\}_{n=1}^k) \}_{j=1}^{T_c}\), where \(X_q \in \mathbb{R}^{H \times W}\) refers to the query image and \(\{X^n_s\}_{n=1}^k\) is its corresponding \(k\) normal images (support images). The PEIM is trained to extract prior information from \(\{X^n_s\}_{n=1}^k\) 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.
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, 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 \( X_s \in \mathbb{R}^{C \times H \times W} \) and the query sample \( X_q \in \mathbb{R}^{C \times H \times W} \) of the target category, the teacher extracts features \( \{ F_{s_i}^T \}_{i=1}^3 \in \mathbb{R}^{C_i \times H_i \times W_i} \) based on \( X_s \) from the first three stages and the student gives corresponding representations \( \{ F_{s_i}^S \}_{i=1}^3 \), where \( C_i, H_i \) and \( W_i \) are the channel, height, and width at the \( i^{th} \) stage, respectively. Then a knowledge distillation loss is used to enforce the feature consistency between them:

\[
\mathcal{L}_{KD}(F_{s_i}^T, F_{s_i}^S) = 1 - \frac{\text{flat}(F_{s_i}^T)}{\|\text{flat}(F_{s_i}^T)\|_2} \cdot \frac{\text{flat}(F_{s_i}^S)}{\|\text{flat}(F_{s_i}^S)\|_2},
\]

where \( \text{flat}(\cdot): \mathbb{R}^{C_i \times H_i \times W_i} \rightarrow \mathbb{R}^{C_i H_i W_i} \) is the flatten function and \( \| \cdot \| \) means the \( l_2 \) norm. In inference, the pixel-wise similarity between \( \{ F_{q_i}^T, F_{q_i}^S \}_{i=1}^3 \) 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 $F_s^{T_i} = \{(F_s^{T_i})^m \in \mathbb{R}^{C_i \times H_i \times W_i}\}$ 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 $3 \times 3$ and a reshape operation on it to produce feature $\bar{F}_s^{T_i} = \{(\bar{F}_s^{T_i})^l \in \mathbb{R}^{H_i \times W_i}\}_{l=1}^L$. 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 $P_s^{T_i} = \{(P_s^{T_i})^l \in \mathbb{R}^{C_i}\}_{l=1}^L$. The whole process can be formulated as follows:

$$
(P_s^{T_i})^l = \frac{H_i W_i}{\sum_{m=1}^L \sum_{h,w} (\bar{F}_s^{T_i})^l,m \cdot (F_s^{T_i})^m},
$$

(2)

where $(h,w)$ indicates the spatial index. We use the orthogonal loss $L_{orth}$ 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 $C_q^i = \{(C_q^i)^l \in \mathbb{R}^{H_i \times W_i}\}_{l=1}^L$ between each prototype $(P_s^{T_i})^l$ and each location $(h,w)$ on the query feature $F_q^S_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ as follows:

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

(3)

Then, for each position, a prototype $(P_s^{T_i})_{h,w}$ with the largest similarity is selected to form the guide feature $F_G^S_i \in \mathbb{R}^{C_i \times H_i \times W_i}$, where $l_{h,w} = \text{argmax}_l (C_q^i)_{h,w}$. We also add up similarity information in $C_q^i$ across all prototypes to get the probability map $F_P^S_i \in \mathbb{R}^{H_i \times W_i}$. 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 $F_q^{S_{i+1}}$:

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

(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 $R_{a,b}^{T_i,S_i} \in \mathbb{R}^{H_i W_i \times H_i W_i}$ between two features $F_a^{T_i}, F_b^{S_i} \in \mathbb{R}^{C_i \times H_i \times W_i}$:

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

(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 $F^T_{q_i}$ and $g_i(F^S_{q_i})$ at $i^{th}$ stage, where $g_i(\cdot)$ is the projection head [1] for transformation, we select the correspondence between teacher representation $F^T_{s_i}$ of the support image and $F^T_{q_i}$ as an anchor correlation, denoted as $R^T_{s,q_i}$. $R^T_{s_i}F^S_{q_i}$ and $R^T_{s,q_i}$ are treated as positive-negative pairs, where $q^-$ belongs to a different category. The contrastive distillation loss with the temperature coefficient $\tau$ is formulated as:

$$L_{NCE}(F^T_{s_i}, F^T_{q_i}, F^S_{q_i}) = - R^T_{s,q_i} \cdot R^T_{s_i} F^S_{q_i} / \tau + \log(\sum_{q^-} e^{R^T_{s,q_i} \cdot R^T_{s_i} F^S_{q_i} / \tau})$$

(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:

$$\mathcal{L} = \sum_{i=1}^{3} [\mathcal{L}_{KD}(F^T_{q_i}, F^S_{q_i}) + \frac{\lambda_1}{k} \sum_{n=1}^{k} \mathcal{L}_{NCE}(F^T_{s_n}, F^T_{q_i}, F^S_{q_i}) + \lambda_2 \cdot L_{orth}(P^T_{s_i})],$$

(7)

where $\lambda_1$ and $\lambda_2$ 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 $256 \times 256$ 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 $\tau$ is 0.05. We employ the default settings in [3] 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 [3] and all experiments are conducted on a Nvidia Tesla V100 GPU.
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.
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>MVTec AD→MPDD</th>
<th>MPDD→MVTec AD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k=2</td>
<td>k=4</td>
</tr>
<tr>
<td>TDG [31]</td>
<td>54.3</td>
<td>59.1</td>
</tr>
<tr>
<td>DifferNet [28]</td>
<td>57.6</td>
<td>62.2</td>
</tr>
<tr>
<td>RegAD [15]</td>
<td>60.1</td>
<td>63.4</td>
</tr>
<tr>
<td>Ours</td>
<td>62.1</td>
<td>65.8</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>PEIM CDS k=2 k=4</th>
<th>k MVTEc AD VisA</th>
<th>N k=2 k=4</th>
<th>L k=2</th>
</tr>
</thead>
<tbody>
<tr>
<td>√ 75.5 76.9 √</td>
<td>2 90.2</td>
<td>0 75.5</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>8 95.3</td>
<td>5 81.5</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>32 95.9</td>
<td>10 85.7</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>64 96.8</td>
<td>14 90.2</td>
<td>100</td>
</tr>
<tr>
<td>(a) Key components on MVTec AD.</td>
<td>(b) Number of shots.</td>
<td>(c) Training categories.</td>
<td>(d) Prototypes.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Backbone MVTec AD VisA MPDD</th>
<th>Operation MVTec AD VisA MPDD</th>
<th>w/o rotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>RegAD [15] 85.7 75.3 63.4</td>
<td>Generation [11] 82.7 77.9 62.4</td>
<td>87.2 81.1 63.9</td>
</tr>
<tr>
<td>ResNet18 86.7 77.3 63.8</td>
<td>Integration [11] 88.4 82.2 65.2</td>
<td></td>
</tr>
<tr>
<td>ResNet34 87.6 80.2 64.5</td>
<td>Ours 90.2 83.4 66.6</td>
<td></td>
</tr>
<tr>
<td>ResNet50 88.9 82.1 65.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WideResNet50 90.2 83.4 66.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(e) Different teacher backbones.</td>
<td>(f) Prototype generation and integration.</td>
<td></td>
</tr>
</tbody>
</table>
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%↑ on MVTec AD, 2.3%↑ on VisA, and 2.7%↑ 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


