Co-Progression Knowledge Distillation with Knowledge Prototype for Industrial Anomaly Detection

Bokang Yang, Zhe Zhang, Jie Ma*

Huazhong University of Science and Technology
{yangbokang, zhangzhe1997, majie}@hust.edu.cn

Abstract

Unsupervised anomaly detection has emerged as a powerful technique for identifying abnormal patterns in images without relying on pre-labeled defective samples. Many unsupervised methods use pre-trained feature extractors from large datasets, with knowledge distillation between teacher and student models being a leading technique. However, due to the similar structures of teacher and student, these methods face challenges like excessive specialization and inadequate generalization, reducing detection performance. In this paper, we introduce a Co-Progression Knowledge Distillation (CPKD) framework, enabling bidirectional learning between teacher and student models. This innovative framework enables concurrent evolution of both models, fostering mutual improvement and enhanced adaptability. To maintain system stability and prevent overspecialization, we introduce a knowledge prototype as a regulatory mechanism for the teacher's learning process. Our method effectively addresses key challenges in anomaly detection, including insufficient learning and overadaptation, by striking a balance between acquiring new knowledge and preserving core competencies. We demonstrate significant improvements in detection accuracy, achieving SOTA performance on the MVTec dataset.

Introduction

Unsupervised anomaly detection is a critical task in computer vision, with applications in industrial quality control(Roth et al. 2022) and medical imaging(Huang et al. 2024). The goal is to identify abnormal patterns in images without relying on pre-labeled defective samples. This approach has gained significant traction in domains where data collection is challenging, such as industrial defect detection and medical imaging. By learning from normal samples, these methods demonstrate remarkable generalization performance for unknown or newly emerged anomalies, making them particularly effective in real-world applications.

Recent advancements in unsupervised anomaly detection have leveraged pre-trained feature extractors derived from large datasets(Bae, Lee, and Kim 2023; McIntosh and Albu 2023). These sophisticated extractors demonstrate remarkable capabilities in distinguishing between normal and abnormal patterns, making them invaluable tools for anomaly detection. The effectiveness of these extractors lies in their ability to produce significantly different outputs for normal and abnormal samples, thereby facilitating the identification of anomalies.

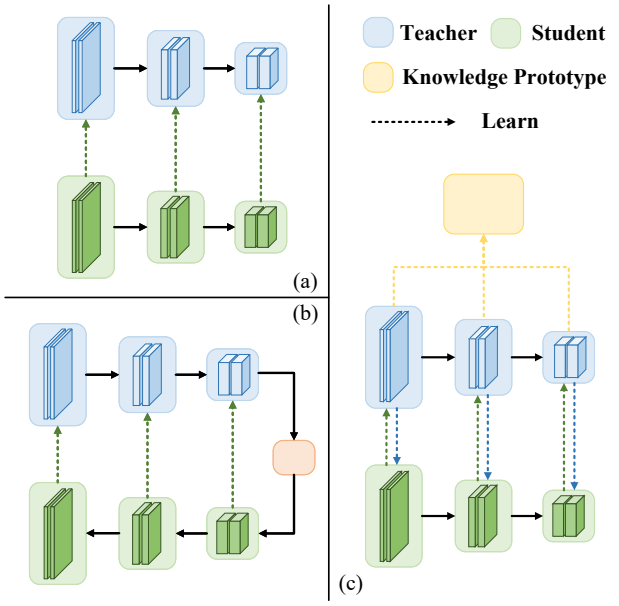

Figure 1: The differences between (a) vanilla knowledge distillation (b) reverse distillation and (c) Co-Progression Knowledge Distillation with Knowledge Prototype. Notice that our paradigm can be used in both forward and reverse distillation.

Knowledge distillation(Gou et al. 2021), a technique employing teacher and student models, has emerged as a leading approach in this field. In this paradigm, the student model is trained to emulate the teacher's output for normal samples. During the inference phase, both models process the input concurrently. When presented with abnormal samples, the student model generates output that notably diverges from the teacher's, allowing anomalies to be detected by quantifying the discrepancy between their respective outputs. Multi-student knowledge distillation (Bergmann et al. 2020) utilizes multiple student models to capture di-

*Corresponding author.

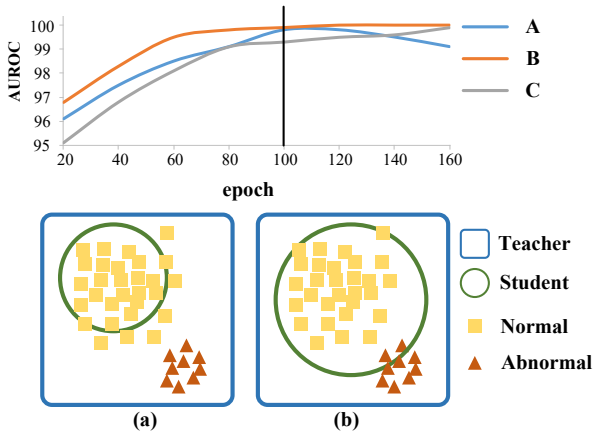

Figure 2: The influence of excessive specialization and inadequate generalization in vanilla knowledge distillation. As shown in the top fig, with the train epoch increasing,the AUROC of class A growth first, but after 100 epochs, it starts to decrease, which represents the excessive specialization. But class C's AUROC keeps increasing, which represents the inadequate generalization. (a)Inadequate generalization: student fails to adequately learn the teacher's response to normal patterns. (b) Excessive specialization: student unintentionally learns abnormal features.

verse aspects of the teacher's knowledge. Feature pyramid-based knowledge distillation (Wang et al. 2021) leverages hierarchical feature representations for more comprehensive knowledge transfer. Reverse distillation (Deng and Li 2022) using encoder-decoder structures inverts the traditional knowledge flow, allowing for unique insights. Despite these advancements, challenges persist due to the structural similarities between teacher and student models. Two primary issues have been identified: Excessive specialization: The student model may inadvertently learn to recognize abnormal features, compromising its ability to detect genuine anomalies. Inadequate generalization: The student model may struggle to fully capture the teacher's response to normal patterns, leading to reduced detection accuracy. Both of these challenges significantly impact the overall performance of anomaly detection systems, necessitating further research and innovation in this domain.

To address these drawbacks, we've uncovered a subtle yet powerful bidirectional learning mechanism in teacher-student interactions. This mechanism manifests not only in students absorbing knowledge from teachers but also in teachers gaining new insights from students' unique perspectives. We term this phenomenon teacher-student Co-Progression (CP). Introducing this concept into knowledge distillation, we've pioneered innovative model training methodologies applicable to both vanilla forward knowledge distillation and reverse distillation.

In traditional model training, teacher models are often viewed as static knowledge repositories(Rudolph et al. 2023). However, we've broken this paradigm, endowing teacher models with the ability to learn and adapt. In this novel training paradigm, teacher models can dynamically adjust their parameters based on interactions with student models. This bidirectional learning mechanism offers significant advantages, accelerating the convergence of teacher and student models in handling routine tasks, enhancing training efficiency, and mitigating potential underfitting in student models. Moreover, it allows the teacher model to refine its knowledge base continuously, adapting to new information and insights from students. This dynamic learning process fosters mutual improvement between teacher and student models, making them gengrate similar representations of normal patterns while maintaining their unique characteristics.

However, the extent to which teachers learn from students must be carefully controlled. To address this challenge, we've introduced the concept of a Knowledge Prototype(KP). This prototype, constructed from another frozen teaching monitor model trained on normal samples, ensures that the teacher model doesn't deviate from its core objectives while absorbing new knowledge. This regulatory mechanism prevents potential "degradation" of teacher models, safeguarding the development of the entire learning ecosystem and preventing overfitting in student models.

Our approach opens up new avenues for model training, potentially leading to more efficient, adaptable, and robust systems. By mimicking the dynamic nature of human learning, we aim to create knowledge distillation that can continuously evolve and improve through interaction, while maintaining their fundamental knowledge base. Our main contributions are summarized as follows:

- We introduce a Co-Progression Knowledge Distillation paradigm, enabling bidirectional learning between teacher and student models. This accelerates convergence, improves efficiency, and enhances detection of diverse anomalies, addressing underfitting issues.
- We propose a Knowledge Prototype framework, developing a regulatory paradigm to control the teacher model's learning process. This maintains system stability and effectiveness, addressing overfitting issues.
- We demonstrate significant improvements in detection accuracy, achieving state-of-the-art performance on the MVTec AD dataset and competitive results on the VisA dataset.

Related Work

Unsupervised image anomaly detection techniques have made significant progress, particularly with the introduction of the MVTec AD dataset(Bergmann et al. 2019). This dataset, consisting of high-resolution images of industrial products with and without defects, has become a benchmark for evaluating anomaly detection algorithms. Current unsupervised methods fall into two main categories: reconstruction-based and feature embedding-based approaches.

Reconstruction-based Methods: These methods employ generative models to reconstruct input images, having been trained exclusively on normal images. They of-

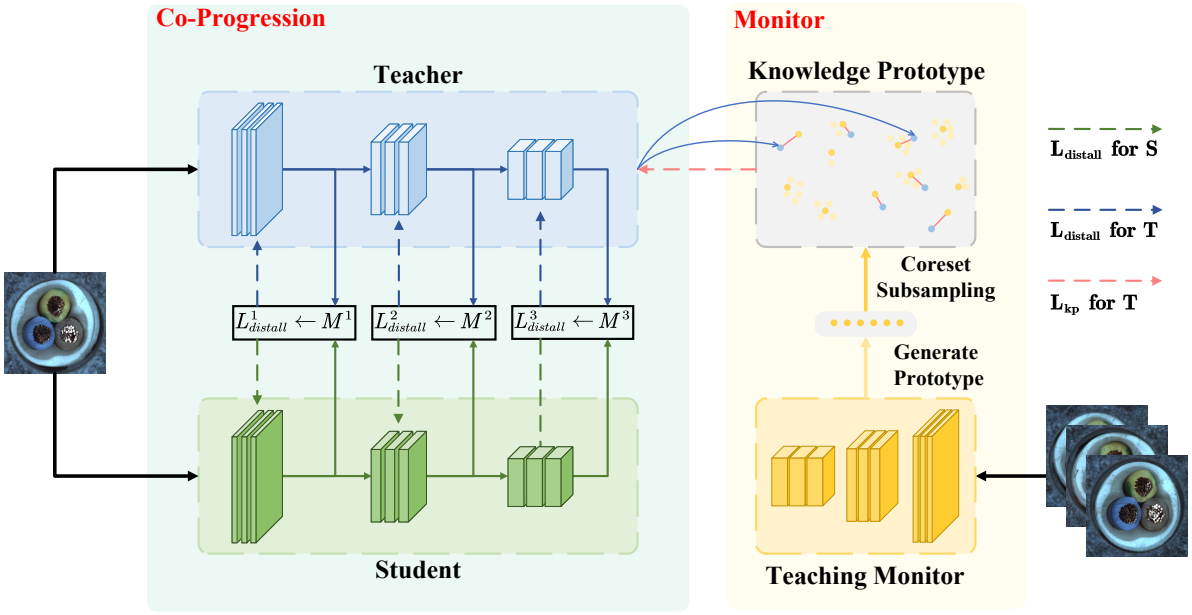

Figure 3: The pipeline of the proposed Co-Progression with Knowledge Prototype framework. Notice we only show the forward KD here, but the structure of reverse KD is the same. The teaching monitor model TM generates a Knowledge Prototype $\mathcal{K}$, which is used to guide the teacher model T during the Co-Progression knowledge distillation process. The student model S learns from the teacher model T and updates based on the knowledge distillation loss. The teacher model T updates based on the knowledge prototype loss and knowledge distillation loss.

ten fail to accurately reconstruct anomalous regions when presented with defective images, thus highlighting anomalies by comparing original and reconstructed images. Early approaches(Zavrtanik, Kristan, and Skočaj 2021; Schlüter et al. 2022) used autoencoder architectures to learn compressed representations of images and attempt to reconstruct them. These simple yet effective models laid the groundwork for more advanced techniques. Generative Adversarial Networks (GAN)(Yan et al. 2021; Liang et al. 2023) generate realistic images and aid in anomaly detection through the discriminator network. The adversarial training process captures fine-grained details of normal samples, making anomalies more apparent during reconstruction. Diffusion Models(Wyatt et al. 2022; Zhang et al. 2023), which generate images through a gradual denoising process, have recently gained attention for their excellent performance in anomaly detection. They offer a unique approach to image generation, potentially capturing nuanced aspects of normal data distribution(Ho, Jain, and Abbeel 2020). Despite their success, reconstruction-based methods face challenges, notably their occasional ability to reconstruct anomalous regions well due to similarities between anomalous and normal areas, which can reduce detection accuracy(Duan et al. 2023; Lu et al. 2023). This issue is particularly significant when anomalies are subtle or structurally similar to normal patterns.

Feature Embedding-based Methods: These methods use pre-trained models to extract high-level semantic features from images, leveraging deep neural networks trained on large-scale datasets. The key idea is that anomalies appear as deviations in the feature space of well-trained models. One-Class Classification(Yi and Yoon 2020). Early methods use Support Vector Data Description (SVDD) aim to learn a compact representation of normal data in the feature space, with deviations treated as anomalies. Distribution Mapping(Rudolph, Wandt, and Rosenhahn 2021; Rudolph et al. 2022) use normalizing flows(Kirichenko, Izmailov, and Wilson 2020) to map features of normal images to a simple distribution, calculating anomaly scores based on distribution probabilities, allowing for flexible modeling of normal data distributions. Memory Bank(Defard et al. 2021; Zou et al. 2022), by storing features of all normal samples and detecting anomalies through feature distance calculations, this method posits that normal samples have similar feature representations, while anomalies deviate significantly. Knowledge Distillation involves detecting anomalies through feature differences between teacher and student model outputs. Multi-student(Bergmann et al. 2020) training structures provide stable differences and(Wang et al. 2021) introduction of feature pyramid effectively capturing anomalies of varying sizes. Concepts like reverse distillation(Deng and Li 2022) and its improvements(Tien et al. 2023) have significantly enhanced detection performance by inverting traditional knowledge transfer processes. This approach has shown promising results, with the teacher model's output serving as a reference for detecting anomalies in the student model's output. However, these methods face challenges due to the similarity in structure between teacher and student models, leading to excessive specialization and inadequate

Algorithm 1: Co-Progression Knowledge Distillation with Knowledge Prototype for Anomaly Detection

Input: Normal samples dataset $\mathcal{D}$, Pre-trained teacher model T, Randomly initialized student model S, Frozen teaching monitor model TM
Parameter: learning rates η_S and η_T , number of epochs E
Output: trained models T, S and knowledge Prototype $\mathcal{K}$

for $I_k \in \mathcal{D}$ **do**
  $\Phi_k = TM(I_k)$
  $V \leftarrow V \cup \text{Unfold}(\Phi_k)$
end for
$\mathcal{K} \leftarrow \text{Coreset}(V, M)$
for epoch $e = 1$ to E **do**
  for $I_k \in \mathcal{D}$ **do**
    $O_T = T(I_k)$
    $O_S = S(I_k)$
    $L_{kd} = Sim\_Cos(O_S, O_T)$
    $S = S - \eta_S \nabla L_{kd}$
    $O_S = S(I_k)$
    $L_{kd} = Sim\_Cos(O_S, O_T)$
    $L_{kp} = KPLoss(\mathcal{K}, O_T)$
    $L_T = \alpha \cdot L_{kp} + (1 - \alpha) \cdot L_{kd}$
    $T = T - \eta_T \cdot \nabla L_T$
  end for
end for

generalization, reducing detection performance.

Method

The proposed Co-Progression with Knowledge Prototype(CPKP) framework, as illustrated, integrates three key components: a pre-trained teacher model T, a student model S, and a teaching monitor model TM that generates a knowledge prototype $\mathcal{K}$. The training process unfolds in two stages: first, TM creates $\mathcal{K}$, establishing a foundation for learning; then, co-progression knowledge distillation occurs. In this second stage, S learns from T's multi-level outputs, while T updates based on both its interaction with the evolving S and its alignment with $\mathcal{K}$. This innovative, bidirectional approach not only enhances the student's learning efficiency but also allows the teacher to adapt dynamically, striking a balance between maintaining core knowledge and accommodating new information. Consequently, the framework achieves improved adaptability and generalization capabilities while preserving overall system stability.

Knowledge Prototype Generation

The knowledge prototype $\mathcal{K} = \{k_1, k_2, ..., k_m\}$ is a crucial component of our co-progression framework, acting as a stabilizing force and repository of essential domain knowledge. $\mathcal{K}$ serves as a reference point, ensuring that the teacher model T maintains consistency with established normal patterns during the co-progression knowledge distillation process. The process begins with the teaching monitor model TM, typically a pre-trained neural network, which processes each image $I_i \in \mathcal{D}$ to extract high-dimensional feature maps. This can be represented as:

$$\Phi_i^l = TM^l(I_i) \in \mathbb{R}^{c_l \times h_l \times w_l} \tag{1}$$

where Φ_i^l is the feature map for image I_i with block number $l \in 1, 2, 3$, and c_l, h_l, and w_l represent the channel, height, and width, respectively. To utilize multi-layer feature maps simultaneously, we upsample deeper layers and aggregate them through an average pooling layer:

$$\Phi_i = \text{AP}(\Phi_i^1, \text{upsample}(\Phi_i^2), \text{upsample}(\Phi_i^3)) \tag{2}$$

To create patch-level prototypes, which retain nearby information crucial for anomaly detection, we use the unfold operation:

$$P_i = \text{Unfold}(\Phi_i, k, s, p) \tag{3}$$

where $P_i \in \mathbb{R}^{(c \cdot k \cdot k) \times N}$,with kernel size k, stride s, and padding p. The number of patches N can be calculated as:

$$N = \left\lfloor \frac{h + 2p - k}{s} + 1 \right\rfloor \cdot \left\lfloor \frac{w + 2p - k}{s} + 1 \right\rfloor \tag{4}$$

We then reshape the unfolded patches into a set of feature vectors and aggregate them:

$$V = \bigcup_{i=1}^{n} V_i = \{\phi_{1,1}, ..., \phi_{1,N}, \phi_{2,1}, ..., \phi_{n,1}, ..., \phi_{n,N}\} \tag{5}$$

This step is essentially a transpose operation on P_i. To reduce computation time and memory usage, we adopt coreset subsampling(Sener and Savarese 2017):

$$\mathcal{K} = \text{Coreset}(V, M) \tag{6}$$

where M is the desired number of samples after subsampling, and $M \ll n \cdot N$. We initialize $\mathcal{K}$ with a random point from V. For each point $\phi \in V$, we compute its distance to the nearest point in $\mathcal{K}$:

$$\mathcal{K} = \mathcal{K} \cup \arg\max_{\phi \in V} \min_{\psi \in \mathcal{K}} |\phi - \psi|_2 \tag{7}$$

This process is repeated until $|\mathcal{K}| = M$. The entire process can be summarized as a composition of functions:

$$\mathcal{K} = \text{Coreset} \circ \text{Unfold} \circ \text{Aggregate} \circ TM(\mathcal{D}) \tag{8}$$

To utilize the knowledge prototype in the learning process, we incorporate a prototype loss term in the teacher model's loss function:

$$L_{kp}(T(I_i), \mathcal{K}) = \min_{k \in \mathcal{K}} |f_{agg}(T(I_i)) - k|_2 \tag{9}$$

where $f_{agg}(\cdot)$ represents the Unfold and Aggregate operations.

Co-Progression Knowledge Distillation

In the CPKD process, the teacher model T and student model S interact in a bidirectional manner, with S learning from T's multi-level outputs and T adapting based on its interactions with S and alignment with the knowledge prototype $\mathcal{K}$. This dynamic learning mechanism accelerates convergence, enhances efficiency, and improves detection

performance, addressing both underfitting and overfitting issues. We use $T^l(I_i)_{hw}$ and $S^l(I_i)_{hw}$ to denote the feature map outputs of the teacher and student models at position (h, w) in the feature map l. The bidirectional learning process is governed by the following loss functions:

$$M^l(h,w) = 1 - \frac{T^l(I_i)hw^T \cdot S^l(I_i)hw}{|T^l(I_i)hw|\,|S^l(I_i)hw|} \tag{10}$$

where $M^l(h, w)$ represents the cosine similarity between the teacher and student model outputs at position (h, w) in the feature map l. The overall similarity score is calculated as the average cosine similarity(Salehi et al. 2021) across all positions in the feature map:

$$\mathcal{L}_{kd} = \sum_{l=1}^{L} \left\{ \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} M^l(h,w) \right\} \tag{11}$$

In our training process, we first optimize the student model S using the $\mathcal{L}_{kd}$. We then update the teacher model T based on the student's output and alignment with the knowledge prototype $\mathcal{K}$.

$$\mathcal{L}_{total} = \alpha \cdot L_{kp} + (1-\alpha) \cdot L_{kd} \tag{12}$$

where α is hyperparameter controlling the relative importance of the knowledge prototype loss L_{kp} and knowledge distillation loss L_{kd}, respectively. The teacher model T is updated using the total loss $\mathcal{L}_{total}$, ensuring that it maintains alignment with the knowledge prototype while transferring knowledge to the student model S.

Inference

In the inference stage, the teacher T and student S models process the input image simultaneously. The student model's output is compared with the teacher model's output, with anomalies detected by calculating the deviation between the two. The anomaly score is calculated as the average cosine similarity across all positions in the feature map:

$$\mathcal{A}_{kd} = \sum_{l=1}^{L} \left\{ \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} M^l(h,w) \right\} \tag{13}$$

The distance between the teacher and knowledge prototype is also calculated to determine the deviation from normal patterns:

$$\mathcal{A}_{kp} = \min_{k \in \mathcal{K}} |f_{agg}(T(I_i)) - k|_2 \tag{14}$$

The final anomaly score is computed as a weighted sum of the knowledge distillation anomaly score $\mathcal{A}_{kd}$ and the knowledge prototype anomaly score $\mathcal{A}_{kp}$:

$$\mathcal{A}_{total} = \gamma \cdot \mathcal{A}_{kd} + (1-\gamma) \cdot \mathcal{A}_{kp} \tag{15}$$

Experiments

Dataset

MVTec (Bergmann et al. 2019) is widely used benchmark for anomaly detection. The dataset consists of 15 object categories, with 5 classes of textures and 10 classes of objects, each containing normal and defective samples. Each class has multiple defective types, with varying degrees of difficulty. The masks for defective regions are provided, enabling quantitative evaluation of detection performance.**VisA** (Zou et al. 2022) is a challenging dataset for anomaly detection, containing 12 classes of industrial products with normal and defective samples. The dataset is characterized by complex textures and subtle anomalies, making it a suitable benchmark for evaluating detection performance.

Experiments details

Experiments settings The size of all input images is 256x256, and use the mean and variance of ImageNet(Deng et al. 2009) for normalization. We use WideResNet50(Zagoruyko and Komodakis 2016) as backbone. For forward distillation, we use the same structure as STPM(Wang et al. 2021), and for reverse distillation, we use the same structure as RD(Deng and Li 2022). The model S are trained using the Adam optimizer with a learning rate of $1e-4$, and the learning rate of T is $1e-6$. The batch size is set to 8, and the number of epochs is 300. The hyperparameters α and γ are set to 0.999 and 0.5, respectively. The coresetsub sampling percentage is set to 1%. The experiments are conducted on a single NVIDIA RTX3080Ti GPU use Pytorch(Paszke et al. 2019).Each experiment is repeated five times, and the average results are reported.

Baselines We compare our proposed CPKP framework with several state-of-the-art anomaly detection methods, including PatchCore(Roth et al. 2022), SimpleNet(Liu et al. 2023), GLAD(Yao et al. 2024), FD(Wang et al. 2021), RD(Deng and Li 2022) and RD++(Tien et al. 2023). FD is a forward knowledge distillation method, RD is a reverse knowledge distillation method, and RD++ is an improved version of RD.

Evaluation Metrics Following the previous work (Bergmann et al. 2019; Zou et al. 2022). We use the Area Under the Receiver Operating Characteristic curve (AUC-ROC) as image-level performance evaluation metrics. AUC-ROC measures the trade-off between true positive rate and false positive rate. And localization performance is evaluated using AUROC at pixel-level.

Results analyses

MVTec As shown in Table 1, our proposed CPKP achieves the highest average image-level AUC-ROC of 99.55%, surpassing other state-of-the-art methods. CPKP consistently performs well across various object categories, indicating robust generalization capabilities. Notably, it excels in specific categories such as Carpet and Bottle, often matching or outperforming existing methods. The CPKP approach shows clear improvements over its base methods (FD and RD), validating the effectiveness of the Co-Progression. Furthermore, CPKP maintains a balanced performance between image-level and pixel-level detection, suggesting enhanced anomaly localization abilities, with an average pixel-level AUC-ROC of 98.01%.It is worth noting that CPKP(RD) achieves the best performance on most categories, indicating the robustness and stability of the proposed framework.

VisA The proposed CPKP method demonstrates competitive performance on the VisA dataset, as evidenced by

	PatchCore	SimpleNet	GLAD	FD	RD	RD++	CPKP(FD)	CPKP(RD)
Carpet	98.7/99.0	99.7/98.2	99.0/98.5	99.1/99.0	98.9/98.9	**100**/99.2	**100**/99.0	**100/99.3**
Grid	98.2/98.7	99.9/98.8	**100/99.6**	99.1/99.0	100/99.3	**100**/99.3	99.9/98.7	**100**/99.1
Leather	**100**/99.3	**100**/99.2	**100/99.8**	97.1/99.1	**100**/99.4	**100**/99.4	**100**/99.5	**100**/99.4
Tile	98.7/95.4	98.7/97.0	**100/98.7**	**100**/96.9	99.3/95.6	99.7/96.6	99.6/96.1	99.7/95.9
Wood	99.2/95.0	**99.5**/94.5	99.4/**98.4**	99.4/96.5	99.2/95.3	99.3/95.8	**99.5**/94.7	99.2/95.6
Bottle	**100**/98.6	**100**/98.0	**100**/98.9	**100**/98.8	**100**/98.7	**100**/98.8	**100/99.0**	**100**/98.6
Cable	99.5/**98.4**	**100**/97.6	99.9/98.1	91.4/95.8	95.0/97.4	99.2/98.4	99.4/97.6	99.7/97.8
Capsule	98.1/98.8	97.8/**98.9**	**99.5**/98.5	75.8/98.6	96.3/98.7	99.0/98.8	95.4/97.1	98.2/98.4
Hazelnut	**100**/98.7	99.8/97.9	**100/99.5**	**100**/98.6	99.9/98.9	**100**/99.2	**100**/98.8	**100**/99.3
Metalnut	**100**/98.4	**100/98.8**	**100/98.8**	99.3/97.2	**100**/97.3	**100**/98.1	**100**/97.5	**100**/97.8
Pill	96.6/97.4	**98.6**/98.6	98.1/97.9	94.1/97.6	96.6/98.2	98.4/98.3	95.3/**98.7**	98.1/**98.7**
Screw	98.1/99.4	**98.7**/99.3	96.9/99.1	93.0/98.8	97.0/**99.6**	98.9/**99.7**	95.2/98.9	98.4/99.5
Toothbrush	**100**/98.7	**100**/98.5	**100/99.4**	99.6/99.0	99.5/99.1	100/99.1	99.1/98.5	**100**/99.0
Transistor	**100**/96.3	100/**97.6**	98.3/96.2	96.4/81.9	96.7/92.5	98.5/94.3	97.1/94.2	**100**/94.0
Zipper	99.4/98.8	99.9/**98.9**	98.5/97.9	90.1/98.8	98.5/98.2	98.6/98.8	95.4/98.0	**100**/97.8
Average	99.1/98.06	99.51/98.12	99.31/**98.62**	95.63/97.04	98.46/97.81	99.44/98.25	98.39/97.75	**99.55**/98.01

Table 1: Quantitative results on the MVTec AD dataset. Image AUC-ROC and pixel AUC-ROC. Best results are in bold, and the second results are underlined.

	PatchCore	SimpleNet	GLAD	FD	RD	RD++	CPKP(FD)	CPKP(RD)
Pcb1	96.0/**99.8**	96.9/**99.8**	**99.6**/99.6	93.8/98.1	97.1/99.7	97/99.7	95.2/99.7	**97.6**/99.7
Pcb2	95.1/98.4	99.2/98.8	**100**/98.6	89.3/97.7	97.0/98.0	97.2/**98.9**	93.7/98.3	96.4/98.3
Pcb3	93.0/98.9	97.1/99.2	**99.9**/98.9	84.1/98.1	96.4/**99.3**	96.8/**99.3**	95.1/98.6	97.3/99.1
Pcb4	99.5/98.3	98.9/98.6	**99.9/99.5**	96.7/97.8	99.8/98.3	99.8/98.8	97.4/98.8	99.9/98.7
Macaroni1	90.1/98.5	97.6/99.6	**99.9/99.8**	93.4/98.7	97.3/99.6	94.0/99.7	94.2/99.0	98.2/99.6
Macaroni2	63.4/93.5	83.4/96.4	**98.9/99.8**	83.9/98.7	89.9/99.4	88/99.7	88.9/99	91.9/99.4
Capsules	68.8/96.5	89.5/99.2	**99.1/99.6**	85.2/97.9	89.5/**99.6**	92.1/99.4	86.9/97.9	93.1/99.4
Candle	98.7/**99.2**	96.9/98.6	**99.9**/94.8	96.4/97.1	94.3/98.5	96.4/98.8	96.6/97.8	95.3/98.7
Cashew	97.7/**99.2**	94.8/99.0	98.4/97.0	96.9/99.0	97.6/93.5	97.8/95.5	**99.0**/99.1	97.1/95.8
Chewinggum	99.1/98.9	**100**/98.5	99.6/**99.1**	96.8/98.3	97.6/98.2	96.4/98.4	97.3/98.5	97.7/97.7
Fryum	91.6/95.9	96.6/94.5	99.4/96.9	99.4/95.6	98.4/**97.1**	95.8/96.9	99.3/92.5	**99.5**/96.8
Pipefryum	99.0/99.3	99.2/99.3	98.9/**99.4**	99.1/98.6	96.2/99.3	99.6/99.1	99.2/99.1	**99.9**/99.3
Average	91/98.03	95.84/98.46	**99.46**/98.58	92.92/97.97	95.93/98.38	95.91/**98.68**	95.23/98.19	96.99/98.54

Table 2: Quantitative results on the VisA dataset. Image AUC-ROC and pixel AUC-ROC. Best results are in bold, and the second results are underlined.

the results in Table 2. CPKP(RD) achieves the second-best average image-level AUC-ROC of 96.99% and the third-best pixel-level AUC-ROC of 98.54%, showcasing its robust anomaly detection capabilities. Notably, CPKP(RD) consistently outperforms its FD-based counterpart and often improves upon the base RD method, particularly in categories such as Macaroni2, Capsules, and Pipefryum. While GLAD excels in image-level detection and RD++ in pixel-level detection, CPKP(RD) maintains a balanced performance across both metrics, indicating its versatility. The method achieves top results in specific categories like Pcb1 and Fryum (image-level), and consistently ranks among the top three in many others. The results highlight the effectiveness of the CPKP framework in detecting anomalies in complex objects, underscoring its potential for real-world applications.

Ablation Study

Table 3 presents an ablation study conducted on the MVTec dataset, examining the impact of two key components of our proposed method: Co-Progression (CP) and Knowledge Prototype (KP). The results demonstrate the synergistic effect of combining these components. When applied individually, CP slightly decreases performance (92.54% for FD and 96.41% for RD), while KP shows a modest improvement (97.13% for FD and 98.97% for RD) compared to the baseline methods (95.63% for FD and 98.46% for RD). This indicates that in co-progression without the knowledge prototype, teacher maybe corrupted by student, the model may struggle to maintain normal patterns, leading to a slight decrease in performance. However, the integration of both CP and KP leads to significant performance gains, achieving the highest image-level AUC-ROC scores of 98.39% for FD and 99.55% for RD. This substantial improvement over the baseline and individual component results underscores the effectiveness of our proposed approach. The ablation study

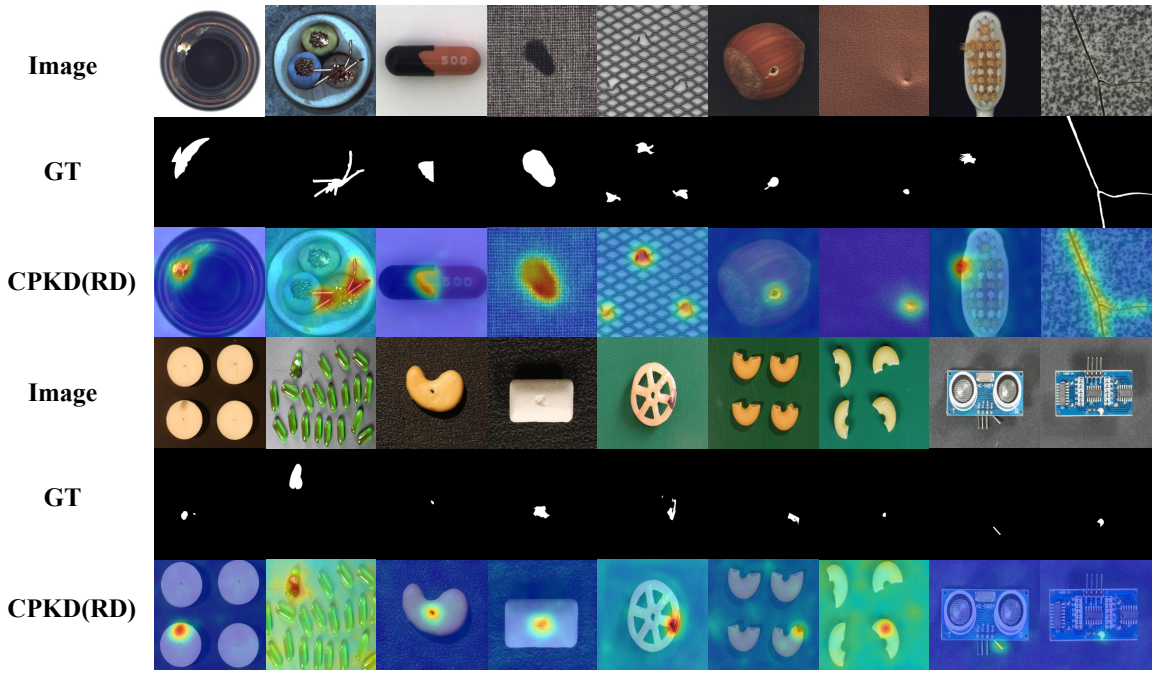

Figure 4: Some visualization results on the MVTec and VisA dataset. The first row shows the input images, the second row shows the ground truth, the third row shows the results of the proposed method.

CP	KP	FD	RD
-	-	95.63/97.04	98.46/97.81
✓	-	92.54/95.13	96.41/96.32
-	✓	97.13/97.21	98.97/97.72
✓	✓	**98.39/97.75**	**99.55/98.01**

Table 3: Ablation study on MVTec dataset image-level and pixel-level AUC-ROC.

α	FD	RD
0.9	97.61/97.13	98.51/97.41
0.99	98.01/97.54	99.20/**98.10**
0.999	**98.39/97.75**	**99.55**/98.01

Table 4: Influence of α on MVTec dataset image-level and pixel-level AUC-ROC.

clearly illustrates that while each component contributes to the overall performance, their combination yields a synergistic effect that surpasses the sum of their individual contributions, highlighting the importance of both Co-Progression and Knowledge Prototype in our method's success.

Influence of hyperparameters

Table 4 illustrates the impact of the hyperparameter α on the performance of our method, evaluated on the MVTec dataset for both image-level and pixel-level anomaly detection. The results demonstrate a clear trend of improved performance as α increases from 0.9 to 0.999 for both FD and RD variants. For the FD variant, the image-level AUC-ROC improves from 97.61% to 98.39%, while the pixel-level AUC-ROC increases from 97.13% to 97.75% as α approaches 0.999. Similarly, the RD variant shows significant improvement, with image-level AUC-ROC rising from 98.51% to 99.55%. Interestingly, for RD, the optimal pixel-level performance (98.10%) is achieved at α = 0.99, with a slight decrease to 98.01% at α = 0.999. These findings suggest that a higher α value, which places more emphasis on the current iteration's prototype in the exponential moving average calculation, generally leads to better performance. The optimal value of $\alpha = 0.999$ for most metrics indicates the importance of giving substantial weight to the prototype while still maintaining a small influence from student.

Conclusion

In this paper, we propose a novel Co-Progression with Knowledge Prototype (CPKP) framework for anomaly detection. CPKP leverages a dynamic, bidirectional learning process between teacher and student models, incorporating a knowledge prototype to maintain core information and enhance generalization capabilities. The proposed method achieves state-of-the-art performance on the MVTec datasets, demonstrating robust anomaly detection capabilities. An ablation study and hyperparameter analysis further validate the effectiveness of CPKP, highlighting the importance of Co-Progression and Knowledge Prototype in enhancing anomaly detection performance.

Acknowledgments

This research was funded by the Interdisciplinary Research Support Program of HUST, grant number 2024JCYJ027.

References

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

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9592–9600.

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

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

Deng, H.; and Li, X. 2022. Anomaly Detection via Reverse Distillation From One-Class Embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9737–9746.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Duan, Y.; Hong, Y.; Niu, L.; and Zhang, L. 2023. Few-shot defect image generation via defect-aware feature manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 571–578.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6): 1789–1819.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, C.; Jiang, A.; Feng, J.; Zhang, Y.; Wang, X.; and Wang, Y. 2024. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11375–11385.

Kirichenko, P.; Izmailov, P.; and Wilson, A. G. 2020. Why normalizing flows fail to detect out-of-distribution data. *Advances in neural information processing systems*, 33: 20578–20589.

Liang, Y.; Zhang, J.; Zhao, S.; Wu, R.; Liu, Y.; and Pan, S. 2023. Omni-frequency channel-selection representations for unsupervised anomaly detection. *IEEE Transactions on Image Processing*.

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

Lu, F.; Yao, X.; Fu, C.-W.; and Jia, J. 2023. Removing anomalies as noises for industrial defect localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16166–16175.

McIntosh, D.; and Albu, A. B. 2023. Inter-realization channels: Unsupervised anomaly detection beyond one-class classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6285–6295.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2022. Towards Total Recall in Industrial Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14318–14328.

Rudolph, M.; Wandt, B.; and Rosenhahn, B. 2021. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1907–1916.

Rudolph, M.; Wehrbein, T.; Rosenhahn, B.; and Wandt, B. 2022. Fully convolutional cross-scale-flows for image-based defect detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1088–1097.

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

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

Schlüter, H. M.; Tan, J.; Hou, B.; and Kainz, B. 2022. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *European Conference on Computer Vision*, 474–489. Springer.

Sener, O.; and Savarese, S. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Tien, T. D.; Nguyen, A. T.; Tran, N. H.; Huy, T. D.; Duong, S. T.; Nguyen, C. D. T.; and Truong, S. Q. H. 2023. Revisiting Reverse Distillation for Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24511–24520.

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

Wyatt, J.; Leach, A.; Schmon, S. M.; and Willcocks, C. G. 2022. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 650–656.

Yan, X.; Zhang, H.; Xu, X.; Hu, X.; and Heng, P.-A. 2021. Learning semantic context from normal samples for unsupervised anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 3110–3118.

Yao, H.; Liu, M.; Wang, H.; Yin, Z.; Yan, Z.; Hong, X.; and Zuo, W. 2024. GLAD: Towards Better Reconstruction with Global and Local Adaptive Diffusion Models for Unsupervised Anomaly Detection. *arXiv preprint arXiv:2406.07487*.

Yi, J.; and Yoon, S. 2020. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian conference on computer vision*.

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

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

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

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