PIRN: PROTOTYPICAL-BASED INTRA-MODAL RECONSTRUCTION WITH NORMALITY COMMUNICATION FOR MULTI-MODAL ANOMALY DETECTION.

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised multimodal anomaly detection (MAD) aims at detecting anomalies by leveraging complementary 2D and 3D data, which plays a crucial role in manufacturing quality control. However, existing MAD methods struggle in *few-shot* scenarios with very limited normal samples, i.e., cross-modal alignment approaches fail to learn reliable correspondences from scarce data, while memory-based methods often identify unseen normal variations as anomalies. To address this gap, we propose **PIRN**, a novel prototype-driven intra-modal reconstruction framework with explicit cross-modal knowledge transfer. Unlike previous work, **PIRN** leverages a compact set of learnable prototypes to capture diverse normal patterns and constrains the reconstruction process to filter out anomalies. It introduces three key innovations: (1) Balanced Prototype Assignment (BPA) via optimal transport ensures uniform prototype usage, preventing codebook collapse, and preserving diverse normal features. (2) Adaptive Prototype Refinement (APR) treats prototypes as adaptive memory, using a gated GRU to update them with each image's normal context; this dynamically expands coverage to unseen normal variations even during testing. (3) Multimodal Normality Communication (MNC) exchanges high-level normal cues between modalities via a gated cross-attention, allowing each modality to assist the other in reconstructing normal features. Extensive experiments on the MVTec 3D-AD and Eyecandies benchmarks show that **PIRN** consistently outperforms state-of-the-art methods in few-shot settings.

1 INTRODUCTION

Multimodal anomaly detection (MAD) Wang et al. (2023); Costanzino et al. (2024); Long et al. (2025b) - the task of identifying defects by jointly inspecting RGB images and 3D point clouds - has become essential for modern manufacturing. Compared with single-modality methods, MAD provides a more complete view of product appearance and can reveal defects that are invisible to either modality alone. Existing MAD methods either rely on cross-modal feature alignment or on memory banks of normal features, but both approaches struggle under few-shot scenarios where only a handful of normal samples per class are available Fang et al. (2023); Tian et al. (2024); Huang et al. (2022). For example, cross-modal alignment approaches such as CFM Costanzino et al. (2024) and LSFA Tu et al. (2025) attempt to learn dense correspondences between RGB and 3D modalities using only normal data. An anomaly is then identified when the features from one modality cannot be predicted by the other. However, with very few normal samples, the learned mapping covers only narrow cross-modal correlations and fails on any unseen correspondence at test time. Memory-bank methods such as M3DM Wang et al. (2023) and SG-DM Chu et al. (2023) store normal feature exemplars and detect anomalies by measuring divergence from all stored samples. With limited normal samples, memory-based models struggle to capture the full range of normal variations, such as pose shifts and texture differences, leading to false positives for mildly deviating test samples. As such, both alignment- and memory-based approaches degrade significantly in data-scarce settings (see Fig. 1 **Left**).

We address these limitations with **PIRN**: **P**rototypical-based **I**ntra-modal **R**econstruction with **N**ormality Communication for few-shot MAD. Rather than overfitting to sparse data via dense cross-modal matching or relying on large memory banks, **PIRN** emphasizes robust **Intra-modal Fea-**

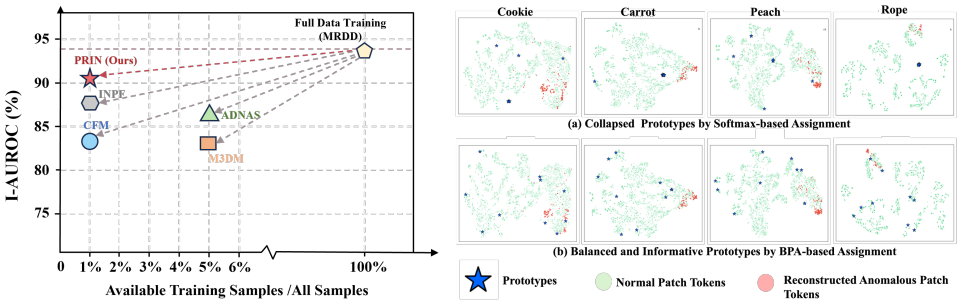


Figure 1: **Left:** Comparison with state-of-the-art methods on the Eyecandies dataset (I-AUROC metric). PIRN achieves superior anomaly detection accuracy using less than 1% of the training data, significantly outperforming existing methods in data-scarce scenarios. **Right:** t-SNE visualization of patch tokens and prototypes in the RGB decoder feature space (MVTec 3D-AD, 10-shot setting). BPA (*bottom*) yields a more uniform prototype distribution over normal features, whereas a softmax assignment (*top*) results in underutilized/collapsed prototypes.

ture Reconstruction using a vector-quantized codebook of discrete normality-aware prototypes Van Den Oord et al. (2017). By reconstructing the features of each modality from a compact codebook, **PIRN** enforces an information bottleneck Alemi et al. (2017); Seo et al. (2023); Zhang et al. (2024b) that retains only essential patterns of normal texture and geometry while ignoring irrelevant details. Consequently, anomalies that cannot be well represented by the prototypes yield large reconstruction errors.

However, naive prototype-based reconstruction presents three major challenges in a few-shot scenario. 1) A naive prototype token assignment scheme (e.g., softmax) suffers from *codebook collapse* Zheng & Vedaldi (2023): a few prototypes can end up encoding common patterns, while others receive fewer updates and remain underutilized. This issue not only downgrades model capacity but also narrows the coverage of "normality" by the prototype codebook. 2) A static prototype codebook learned from scarce training data may not cover all normal variations at test time Zhang et al. (2024a); Wei et al. (2023). In other words, a normal test sample might contain unseen yet still-normal patterns that cannot align with any learned prototype and result in false-positive predictions. 3) Vanilla prototype learning treats each modality in isolation, ignoring the complementary information between texture and geometry Mao et al. (2025). Without effective cross-modal collaboration, subtle defects unique to one modality may go undetected.

We address these challenges with three key innovations built upon the vanilla prototype-based AD framework. **First**, Balanced Prototype Assignment (BPA) formulates patch-to-prototype matching as a balanced optimal transport Peyré & Cuturi (2019) problem, ensuring that each prototype captures a distinct normal pattern. This promotes uniform prototype utilization during patch reconstruction, preventing codebook collapse and enabling the model to represent diverse normal patterns even with limited training examples. As shown in Fig. 1 **Right**, this balanced assignment yields a much more uniform prototype distribution than using softmax.

Second, Adaptive Prototype Refinement (APR) bridges the train–test distribution gap by treating the prototypes as adaptive memory at inference. APR uses a lightweight GRU to update the prototype vectors based on the test image's normal context, without corrupting them with anomaly contexts. This on-the-fly refinement expands the prototypes' coverage to new normal variations that are absent during training. **Third**, we introduce Multi-modal Normality Communication (MNC) that exchanges prototypical normality knowledge across modalities via a two-stage process. The first stage aligns high-level normal concepts encoded by prototypes across modalities through graph refinement. In the second stage, these refined prototypes serve as anchors to guide fine-grained feature reconstruction via cross-attention. As such, this allows each modality to reinforce the other's understanding of normality, enabling more discriminative detection of challenging anomalies (e.g., subtle defects) that might go undetected when each modality is used in isolation.

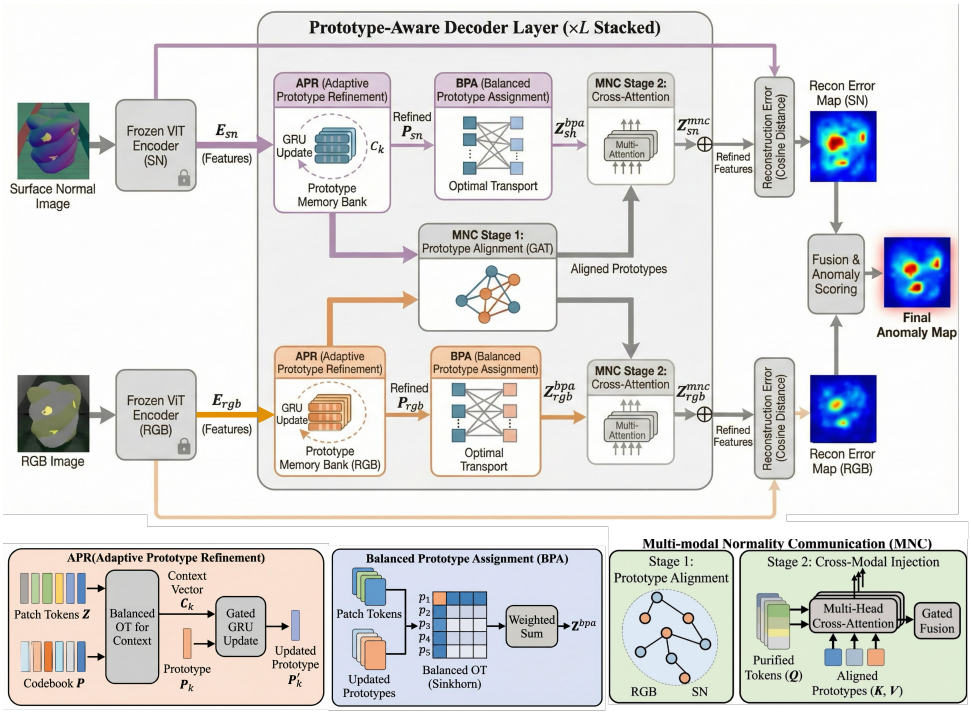


Figure 2: (a) **Overview of PIRN.** Given an RGB image and a surface normal, PIRN uses pretrained frozen encoders to extract features E_{rgb} and E_{sn}. A prototype-aware multi-layer decoder then reconstructs these features into $\mathbf{Z}^{\mathbf{bpa}}$ (intra-modal purified) and $\mathbf{Z}^{\mathbf{mnc}}$ (cross-modal purified), which are used to generate anomaly maps. PIRN introduces three key components: 1) APR for adaptive prototype refinement to capture unseen normal patterns at test time; and 2) BPA for balanced prototype assignment to mitigate codebook collapse; and 3) MNC for cross-modal prototype communication. (b) Details of the three components.

Together, these modules enable our model to learn and communicate normal patterns effectively across modalities, significantly improving anomaly detection in data-scarce settings. Our main contributions are summarized as follows:

- We present **PIRN** – a robust *Prototypical-based Intra-modal Reconstruction with cross-modality Normality Communication* framework for few-shot MAD.
- We introduce BPA to prevent codebook collapse and capture more diverse normal patterns. A lightweight APR module is further proposed to expand the prototypes' coverage to unseen yet normal variations at inference.
- We propose an MNC mechanism that shares normal information across modalities via cross-modal knowledge transfer, enabling each modality to help reconstruct the other's normal features and clearly highlight anomalies.

2 Related Work

2D Anomaly Detection. Many recent 2D anomaly detection (AD) methods constrain normal feature representations by using discrete prototypes to encode "normality." For example, HVQ-Trans Lu et al. (2023) preserves typical normal patterns as a vector-quantized prototype codebook, preventing the "identical shortcut" issue and ensuring anomalies cannot be perfectly reconstructed. Similarly, RLR He et al. (2024) introduces a learnable reference representation to discourage shortcut solutions and explicitly model normal patterns. DPDL Wang et al. (2025) learns multiple Gaussian prototypes

and diffuses normal samples toward these cluster centers, forming a compact normal feature space to exclude anomalies. INP-Former Luo et al. (2025b) extracts intrinsic normal prototypes directly from each test image, eliminating reliance on external memory bank and achieving state-of-the-art performance in 2D AD tasks. Gong et al. (2019) introduces MemAE, a memory-augmented autoencoder that utilizes an explicit memory bank to record prototypical normal patterns, thereby constraining reconstruction to learned normality. Guo et al. (2023) proposes a template-guided approach, utilizing exemplars from the normal training library to guide the hierarchical restoration of input features, detecting anomalies via reconstruction deviations. However, lacking explicit cross-modal interaction, such methods are not directly applicable to MAD tasks.

Multi-modal Anomaly Detection. Existing MAD methods mostly rely on cross-modal alignment or memory banks, with some exploring architecture search and distillation. Cross-modal alignment approaches (e.g., CFM Costanzino et al. (2024), LSFA Tu et al. (2024)) learn to align RGB and 3D features using only normal data, detecting anomalies when one modality's features cannot be predicted by the other. These methods fuse texture and geometry cues effectively but need diverse normal samples to establish reliable cross-modal correspondences. Alternatively, memory-based models such as M3DM Wang et al. (2023) and SG-DM Chu et al. (2023) store normal feature patterns (either fused or modality-specific) and identify deviations as anomalies. Such methods suffer in few-shot settings: any unseen yet normal pattern not shown in the memory can lead to misidentification. Beyond alignment and memory methods, 3D-ADNAS Long et al. (2025b) optimizes feature fusion architectures via neural architecture search.

3 METHOD

3.1 FRAMEWORK OVERVIEW

To the best of our knowledge, **PIRN** (overview in Fig. 2) is the first multi-modal anomaly detection (AD) framework to integrate a vector-quantized prototype codebook into a Vision Transformer (ViT) Dosovitskiy et al. (2020) encoder–decoder architecture. Specifically, for each modality, we learn a compact codebook of K vector-quantized discrete prototypes. These prototypes serve as reference points for typical normal textures and geometries, constraining reconstruction to rely solely on normal information.

Frozen ViT Encoder. We employ two parallel ViT encoders, $\mathcal{E}rgb$ and $\mathcal{E}sn$, which are pre-trained and kept frozen. $\mathcal{E}rgb$ processes the input RGB image, while $\mathcal{E}sn$ processes the corresponding surface-normal map. We extract multi-scale features from a set of intermediate layers of each encoder and aggregate them via element-wise averaging to form a single feature map per modality (denoted E_{rgb} and E_{sn}, each in $\mathcal{R}^{N\times C}$). These aggregated feature maps serve as both the input to the decoder and the target for reconstruction.

Cascaded Prototype-Aware Decoder. The decoder consists of a stack of prototype-aware layers that progressively reconstruct a normal version of input features. Each decoder layer performs three sequential operations. First, Adaptive Prototype Refinement (APR) updates each modality's prototype codebook via a gated recurrent unit (GRU) Chung et al. (2014), enhancing adaptability to the current sample. Next, Balanced Prototype Assignment (BPA) assigns each patch token to the updated prototypes via balanced optimal transport, promoting uniform prototype utilization. Finally, Multi-Modal Normality Communication (MNC) aligns the refined prototypes from both modalities through graph-based attention, and then exchanges high-level normality knowledge between the two modalities.

3.2 BALANCED PROTOTYPE ASSIGNMENT (BPA)

Allowing each token to softly match against all K prototypes can lead to a codebook collapse: some prototypes may eventually become under-utilized, reducing the diversity of normal patterns the codebook can represent. BPA addresses this issue by formulating the token-to-prototype assignment as a balanced *optimal transport* (OT) problem. Instead of using softmax assignment that might over-concentrate on a few prototypes, BPA enforces two crucial properties for a more uniform prototype usage: (1) **patch-to-prototype selectivity**, ensuring each patch token is matched to only a small

subset of prototype codes; and (2) **uniform prototype utilization**, ensuring all prototypes receive a balanced share of patch assignments. Therefore, BPA encourages each prototype to specialize in a distinct normal pattern, yielding a more diverse and representative codebook.

Specifically, let $\mathbf{Z} = \{z_n\}_{n=1}^N$ denote the set of N patch tokens input to a given decoder layer (for the first decoder layer, $\mathbf{Z}$ equals the encoder output E). Let $P = \{p_k\}_{k=1}^K$ denote the prototype vectors of a specific modality's codebook. In practice, before applying BPA we first refine the prototypes using APR (detailed in the next section), which adapts P to the normal context of the input image. This ensures that BPA operates on prototypes already tailored to the current sample.

We then define a cost matrix $C \in \mathcal{R}^{N\times K}$ with entries $C_{nk} = 1 - \frac{z_n \cdot p_k}{\|z_n\|\|p_k\|}$ representing the cosine distance between patch token z_n and prototype p_k. BPA seeks an optimal transport plan $T^* \in \mathcal{R}_{\geq 0}^{N\times K}$ that minimizes the assignment cost under equal-mass constraints:

$$\begin{aligned} T^* \;&=\; \arg\min_{T} \sum_{n=1}^{N}\sum_{k=1}^{K} T_{nk}\, C_{nk} \\ &\text{s.t.} \quad T\,\mathbf{1}_K = \mathbf{a}, \quad T^\top \mathbf{1}_N = \mathbf{b}, \end{aligned} \tag{1}$$

where $\mathbf{a} = \frac{1}{N}\mathbf{1}_N$ and $\mathbf{b} = \frac{1}{K}\mathbf{1}_K$.

This optimal transport formulation yields a balanced soft assignment, avoiding trivial solutions (e.g., all patches assigned to a single prototype) and ensuring full prototype utilization. We solve it using the Sinkhorn algorithm Cuturi (2013) with entropic regularization, which typically converges in a few iterations to the optimal plan T^*. We then use T^* to reconstruct each patch token as a weighted combination of those prototypes.

$$z_n^{bpa} = \sum_{k=1}^{K} T_{nk}^* p_k. \tag{2}$$

This effectively projects the input query tokens $\{z_n\}_{n=1}^N$ onto the prototype space under the learned OT weights T^*.

BPA thus acts as an information bottleneck by reconstructing each patch token using only a limited set of normality-aware prototypes, thereby filtering out anomalous details. Since only normal patterns can be faithfully reconstructed, any anomalous regions in the query input will be poorly reproduced, leading to large reconstruction errors at test time. We refer to $\mathbf{Z^{bpa}} = \{z_n^{bpa}\}$ as *intra-modal purified reconstruction*, since they are derived solely from the normal prototypes of the same modality.

3.3 Adaptive Prototype Refinement (APR)

Our framework relies on a set of prototypes $\{p_k\}_{k=1}^K$ as a compact codebook of normal patterns. A key strength lies in its ability to adaptively adjust these prototypes via a unified refinement procedure applied during both training and testing. This allows the model to capture diverse normal patterns during training and adapt to unseen variations at test time. To achieve this, we introduce Adaptive Prototype Refinement (APR), which dynamically refines the prototypes using the normal context extracted from the current input. Importantly, APR operates on the patch tokens $\mathbf{Z}$ that are extracted from the previous decoder layer (or the encoder's output for the first decoder layer), before any reconstruction is performed in the current decoder layer.

Normal Context Extraction via Optimal Transport. To ensure that only normal patch tokens can contribute to each prototype, we compute an optimal transport alignment between the patch tokens and the prototypes. Similar to Eq. 1, we derive the OT plan Γ^*. This plan associates each prototype p_k with a weighted subset of patch tokens in $\mathbf{Z} = \{z_n\}_{n=1}^N$ that it best represents. We then compute a context vector for prototype p_k as the weighted average of its assigned patch features:

$$c_k \;=\; \frac{\sum_{n=1}^{N} \Gamma_{nk}^*\, z_n}{\sum_{n=1}^{N} \Gamma_{nk}^*}\,. \tag{3}$$

This OT-based context extraction provides robust guidance for prototype refinement. By finding an optimal matching between prototypes and patch tokens, any out-of-distribution (anomalous) patch

will incur a high cost to all prototypes and thus receive a negligible weight. This ensures that each prototype p_k is updated using reliable in-distribution (normal) patches while ignoring anomalous ones, allowing robust refinement even in the presence of minor anomalies.

Gated Prototype Update via GRU. We then update each prototype by incorporating its context vector c_k through a GRU-based gating mechanism. In this update, the original prototype p_k is treated as the hidden state of the GRU and the context c_k as input, producing an updated prototype p'_k. The GRU's gating mechanism dynamically controls the integration of new context, promoting normal features while suppressing anomalous ones during testing. In particular, if the context c_k is unreliable and does not match any existing prototype (e.g., biased by an anomaly), the GRU's update gate will remain largely closed, leaving p_k essentially unchanged. This gating strategy is crucial for maintaining prototype reliability: it allows prototypes to gradually evolve with unseen normal variations at test time without drifting toward anomalies. As such, the model expands its coverage of normal patterns and reduces false positives.

3.4 Multi-Modal Normality Communication (MNC)

To model the complementary cues from texture (RGB) and geometry (surface normals), we introduce a Multi-Modal Normality Communication (MNC) module to exchange normal information between the two branches. The key idea is that each modality can assist the other in understanding normality, thereby better highlighting true anomalies and suppressing false positives. To ensure robust knowledge transfer, MNC exchanges prototype-based normal knowledge between modalities, rather than raw patch features that may contain anomalies during testing. The decoder of each modality is guided to reconstruct features not only from its own prototypes but also from high-level normal patterns of the other modality. MNC operates in two stages: a *prototype alignment* stage and a *cross-modal normality injection* stage.

Stage 1: 2D and 3D Prototype Alignment. We treat all prototypes from both modalities as nodes in a unified graph and perform cross-modal message passing to align them. Specifically, we construct a graph with $2K$ nodes, consisting of K RGB prototypes and K surface-normal prototypes. We connect each prototype to its nearest neighbors in the other modality using KNN in the feature space of L_2-normalized prototypes, and then apply a multi-head Graph Attention Network (GAT) Veličković et al. (2018) to propagate information across these edges. This graph-based refinement pulls the two sets of prototypes into a shared semantic space: prototypes representing similar structures (e.g., a flat surface or an edge) are drawn closer and enriched with complementary context from the other modality. Let P'_{rgb} and P'_{sn} denote the refined RGB and surface-normal prototype sets after this alignment. As a result, the two branches obtain *aligned prototypes* that encode a consistent cross-modal notion of normal texture and geometry. Similar prototype-level alignment strategies Huang et al. (2025); Tang et al. (2023); Pahde et al. (2021) have proven effective in multimodal representation learning.

Stage 2: Cross-Modal Normality Injection. After alignment, the refined prototypes serve as anchors to guide fine-grained feature reconstruction via cross-attention. In this stage, each patch token from one modality will attend to the other modality's refined prototypes to inject any normal information it lacks. To filter out anomalous details in an anomalous test sample, we first purify each modality's patch tokens using its intra-modal information. Specifically, we use the intra-modal purified tokens z_n^{bpa} as an attention mask to reweight the original patch tokens z_n channel-wise. This yields purified tokens $\mathbf{Z}' = \{z_n \cdot \sigma(z_n^{bpa})\}_{n=1}^{N}$, where $\sigma(\cdot)$ is the sigmoid function. These purified tokens $\mathbf{Z}'$ are then used as queries in the cross-modal attention.

For cross-modal knowledge exchange, we employ a cross-attention layer Vaswani et al. (2017) where the refined prototypes of one modality act as keys and values, and the purified patch tokens from the other modality act as queries. Taking the RGB branch as example, let $\mathbf{Z}'_{rgb}$ denote the purified tokens of RGB branch and $\mathbf{P}'_{sn}$ denote the set of stage-1 refined prototypes from the surface normal branch. We compute the cross-attention output as:

$$\mathrm{CA}(\mathbf{Z}'_{rgb}, \mathbf{P}'_{sn}) = \mathrm{SoftMax}\Big(\frac{\mathbf{Z}'_{rgb} W_Q (\mathbf{P}'_{sn} W_K)^\top}{\sqrt{d}}\Big)(\mathbf{P}'_{sn} W_V), \tag{4}$$

where W_Q, W_K, W_V are projection matrices and d is the channel dimension of the Z'_{rgb}.

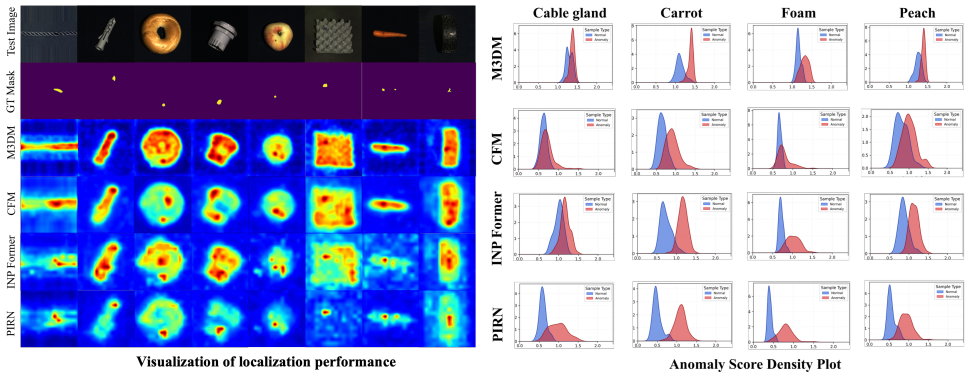


Figure 3: **Left:** Compared to existing MAD methods (10-shot), our anomaly maps are sharper with fewer false positives. **Right:** Comparison of anomaly score distributions for normal and anomalous samples (10-shot, MVTec-3D-AD). **PIRN** shows clearer distribution separation.

To prevent overwhelming the patch features with irrelevant information, we introduce a learnable gating scalar γ to modulate the cross-attention output. Specifically, we add a scaled version of the cross-attention result to the original token representation:

$$\mathbf{Z}^{mnc} = \mathbf{Z}' + g \cdot \mathrm{CA}(\mathbf{Z}', \mathbf{P}'), \qquad g = \tanh(\gamma), \tag{5}$$

where γ is a learnable scalar parameter and $g = \tanh(\gamma)$ serves as a gate on the cross-modal information. This gating mechanism allows the network to control the extent of cross-modal fusion for each layer. By exchanging high-level normality knowledge and injecting it into fine-grained patch tokens, MNC establishes a robust correspondence between modalities at the prototype level. Unlike methods that attempt dense patch-to-patch alignment between modalities (which can be unreliable given limited data), our prototype-centric exchange avoids direct dense mappings and thus offers greater robustness on unseen test samples.

We refer to $\mathbf{Z}^{mnc}$ as *cross-modal purified reconstruction*, as they are obtained using normal prototypes from both modalities. We then fuse $\mathbf{Z}^{bpa}$ and $\mathbf{Z}^{mnc}$ via element-wise summation to produce the final reconstructed features for each modality $\mathbf{Z}^{rec} = \mathbf{Z}^{bpa} + \mathbf{Z}^{mnc}$.

Training and Inference We train **PIRN** end-to-end using an intra-modal feature reconstruction loss (e.g., a soft mining loss Luo et al. (2025a)) that aligns the decoder outputs with the encoder features. In practice, we minimize the cosine distance between the encoder's patch embeddings (E_{rgb} and E_{sn}) and the corresponding reconstructed embeddings in $\mathbf{Z}^{rec}$, across all spatial locations and for both modalities.

At inference time, we compute an anomaly score map by comparing the original encoder features to the reconstructed features at each spatial location. For the i-th patch, the anomaly score is defined as $d_i = 1 - \cos(E_i, \mathbf{Z}_i^{rec})$. This patch-level anomaly map is then upsampled to the input resolution and optionally smoothed with a Gaussian filter. We obtain separate anomaly maps from the RGB and surface-normal branches, which are summed to produce a fused anomaly heatmap. The final image-level anomaly score is taken as the maximum value in this fused heatmap.

4 Experiments

Datasets Setting. We evaluate **PIRN** on the MVTec-3D-AD Bergmann et al. (2022) and Eyecandies Bonfiglioli et al. (2022) datasets under data-scarce conditions by randomly sampling 5, 10 or 50 normal images per class, along with the full-data setting. For each k-shot setting, we repeat the random selection 10 times and report the average performance to mitigate selection bias.

Implementation Details. We adopt a ViT-Base/14 transformer as the backbone encoder for both RGB and surface normal inputs, initialized with DINOv2 Oquab et al. (2023) pre-trained weights and

k-Shot	Method	MVTec-3D-AD			Eyecandies		
		AUROC$_I$	AUROC$_P$	AUPRO	AUROC$_I$	AUROC$_P$	AUPRO
5	BTF Horwitz & Hoshen (2023)	0.671	0.980	0.920	0.652	0.815	0.738
	AST Rudolph et al. (2023)	0.680	0.950	0.903	0.633	0.741	0.691
	M3DM Wang et al. (2023)	0.822	0.984	0.937	0.764	0.871	0.807
	CFM Costanzino et al. (2024)	0.811	0.986	0.949	0.795	0.879	0.801
	3D-ADNAS Long et al. (2025a)	0.826	–	–	0.775	0.875	–
	INP-Former Luo et al. (2025a)	0.851	0.988	0.957	0.859	0.946	0.862
	Ours	**0.890**	**0.990**	**0.960**	**0.895**	**0.955**	**0.887**
10	BTF Horwitz & Hoshen (2023)	0.695	0.983	0.928	0.685	0.834	0.806
	AST Rudolph et al. (2023)	0.689	0.946	0.835	0.671	0.767	0.624
	M3DM Wang et al. (2023)	0.845	0.986	0.943	0.824	0.890	0.812
	CFM Costanzino et al. (2024)	0.845	0.987	0.954	0.838	0.903	0.825
	3D-ADNAS Long et al. (2025a)	0.848	–	–	0.807	0.869	–
	INP-Former Luo et al. (2025a)	0.885	0.989	0.960	0.872	0.947	0.870
	Ours	**0.922**	**0.991**	**0.966**	**0.912**	**0.969**	**0.896**
50	BTF Horwitz & Hoshen (2023)	0.806	0.989	0.947	0.721	0.856	0.824
	AST Rudolph et al. (2023)	0.794	0.974	0.929	0.739	0.862	0.715
	M3DM Wang et al. (2023)	0.907	0.989	0.955	0.836	0.933	0.846
	CFM Costanzino et al. (2024)	0.906	0.991	0.965	0.852	0.926	0.851
	ADNAS Long et al. (2025a)	0.890	–	–	0.868	0.912	–
	INP-Former Luo et al. (2025a)	0.921	0.991	0.965	0.902	0.967	0.892
	Ours	**0.945**	**0.993**	**0.970**	**0.924**	**0.975**	**0.908**
All	BTF Horwitz & Hoshen (2023)	0.865	0.992	0.959	0.740	0.883	0.845
	AST Rudolph et al. (2023)	0.937	0.976	0.944	0.780	0.902	0.744
	M3DM Wang et al. (2023)	0.945	0.992	0.964	0.882	0.977	0.887
	CFM Costanzino et al. (2024)	0.954	0.993	0.971	0.881	0.974	0.887
	3D-ADNAS Long et al. (2025a)	0.951	–	–	0.946	0.970	–
	INP-Former Luo et al. (2025a)	0.952	0.994	0.971	0.934	0.981	0.918
	Ours	**0.963**	**0.994**	**0.973**	**0.948**	**0.983**	**0.923**

Table 1: Comparison of anomaly detection and localization performance on **MVTec-3D-AD** and **Eyecandies** under different training shots.

Modules			Metrics		
BPA	APR	MNC	AUROC$_I$	AUROC$_P$	AUPRO
✗	✗	✗	0.828	0.976	0.952
✗	✓	✓	0.883	0.990	0.956
✓	✗	✓	0.916	0.990	0.961
✓	✓	✗	0.867	0.988	0.947
✓	✓	✓	0.922	0.991	0.966

Table 2: Ablation of different components on **MVTec-3D-AD**.

Method	AUROC$_I$	AUROC$_P$	AUPRO
Softmax Attention	0.832	0.967	0.929
Linear Attention	0.845	0.968	0.931
Sigmoid Attention	0.878	0.976	0.954
Balanced Optimal Transport	0.922	0.991	0.966

Table 3: Ablation of prototype assignment in BPA on **MVTec-3D-AD**.

Method	AUROC$_I$	AUROC$_P$	AUPRO
w/o APR module	0.916	0.990	0.961
Global Averaging	0.915	0.989	0.960
Top-k Averaging	0.921	0.991	0.964
Balanced Optimal Transport	0.922	0.991	0.966

Table 4: Ablation of token aggregation methods in APR on **MVTec-3D-AD**.

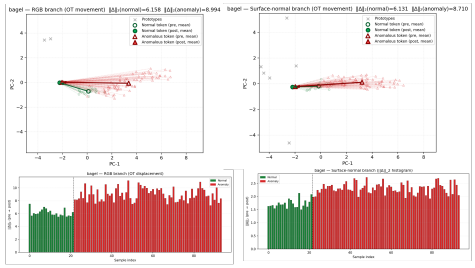
Table 5: Comparisons of per-category anomaly detection performance on MVTec-3D-AD.

Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
I-AUROC											
BTF Horwitz & Hoshen (2023)	0.938	0.765	0.972	0.888	0.960	0.664	0.904	0.929	0.982	0.726	0.865
AST Rudolph et al. (2023)	0.983	0.873	0.976	0.971	0.932	0.885	0.974	0.981	**1.000**	0.797	0.937
M3DM Wang et al. (2023)	0.994	0.909	0.972	0.976	0.960	0.942	0.973	0.899	0.972	0.850	0.945
CFM Costanzino et al. (2024)	0.994	0.888	**0.984**	**0.993**	**0.980**	0.888	0.941	0.943	0.980	**0.953**	0.954
3D-ADNAS Long et al. (2025a)	**0.997**	**1.000**	0.971	0.986	0.966	0.948	0.897	0.873	**1.000**	0.867	0.951
Shape Guided Chu et al. (2023)	0.986	0.894	0.983	0.991	0.976	0.857	0.990	**0.965**	0.990	0.869	0.947
PIRN	0.971	0.973	0.941	0.957	0.975	**0.993**	**0.992**	0.950	0.996	0.880	**0.963**
AUPRO@30%											
BTF Horwitz & Hoshen (2023)	0.976	0.969	0.979	**0.973**	0.933	0.888	0.896	0.912	0.950	0.971	0.959
AST Rudolph et al. (2023)	0.970	0.947	0.981	0.939	0.913	0.906	0.979	0.982	0.889	0.940	0.944
M3DM Wang et al. (2023)	0.970	0.971	0.979	0.950	0.941	0.932	0.977	0.971	0.971	0.975	0.964
CFM Costanzino et al. (2024)	0.979	0.972	0.982	0.945	0.950	0.968	0.980	0.943	0.950	**0.981**	0.971
Shape Guided Chu et al. (2023)	**0.981**	0.973	0.982	0.971	0.962	**0.978**	**0.981**	**0.983**	**0.974**	0.975	**0.976**
PIRN	0.966	**0.978**	**0.983**	0.972	**0.976**	0.971	**0.981**	0.978	**0.974**	0.951	0.973

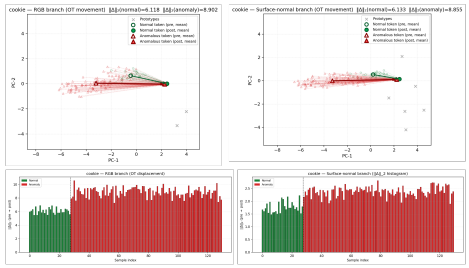
kept **frozen** during training. 3D point clouds are converted into 3-channel surface normal maps and resized to 224×224, as with the RGB images. To obtain robust multi-scale features, we aggregate the patch tokens extracted from the 2-10 layers of the pretrained ViT by element-wise averaging. The decoder is a cascaded architecture of 2 layers for few-shot tasks and 8 layers for all-shot tasks. Model optimization is done using ADAM Kingma & Ba (2014) (learning rate 1×10^{-4}) for 80 epochs in few-shot tasks and 30 epochs in all-shot tasks.

Main Results. As shown in Tab. 1, our method consistently outperforms the best-performing baseline on both *MVTec-3D-AD* and *Eyecandies* across all metrics in varying few-shot settings. Specifically, on *MVTec-3D-AD*, it achieves improvements of 3.9↑ (AUROC$_I$) at 5-shot, 3.7↑ at 10-shot, and 2.4↑ at 50-shot, compared to the strongest baseline. Similarly, on the *Eyecandies* dataset, our approach yields notable gains of 3.6↑ in 5-shot, 4.0↑ in 10-shot, and 2.2↑ in 50-shot, outperforming the best baseline in all metrics. The consistent improvements in few-shot settings validate the effectiveness of our method with extremely limited training data. Notably, **PIRN** also achieves the best performance in the full-shot setting, albeit with a slight margin. On the MVTec-3D-AD full-data performance Tab. 5, PIRN method achieves the highest mean I-AUROC of 0.963, outperforming the other baseline methods.

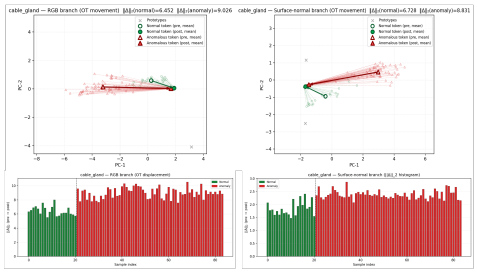
Qualitative Analysis. As shown in Fig. 3 **Left**, our method produces superior anomaly maps compared to baselines. **PIRN** accurately localizes anomalies while suppressing background textures, enabling more *discriminative localization* with fewer false positives. Furthermore, Fig. 3 **Right**



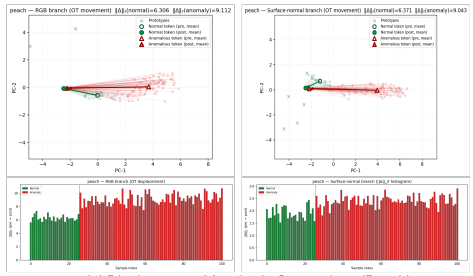
(1) Displacement Magnitude Comparison (Bagel)



(2) Displacement Magnitude Comparison (Cookie)



(3) Displacement Magnitude Comparison (Cable gland)



(4) Displacement Magnitude Comparison (Peach)

Figure 4: Visualization of Feature Displacement via BPA Routing. Green: normal; Red: anomalous.

Table 6: Comparison of different methods based on Real-IAD D3 dataset. The highest value in each row is marked in red, and the second highest value is marked in blue.

Modality	RGB				3D		2D+3D						D^3		RGB + SN	
Model	Cflow		SimpleNet		PointMAE		AST		PointMAE+PatchCore		M3DM		D^3M		PIRN (Ours)	
Metrics	I-AUROC	P-AUROC	I-AUROC	P-AUROC	I-AUROC	P-AUROC	I-AUROC	P-AUROC	I-AUROC	P-AUROC	I-AUROC	P-AUROC	I-AUROC	P-AUROC	I-AUROC	P-AUROC
audio_jack_socket	0.943	0.944	0.973	0.926	0.763	0.655	0.860	0.590	0.926	0.673	0.981	0.699	0.983	0.757	0.950	0.964
common_mode_filter	0.271	0.847	0.717	0.822	0.725	0.687	0.899	0.802	0.523	0.922	0.580	0.934	0.618	0.947	0.826	0.883
connector_housing-female	0.839	0.921	0.795	0.891	0.958	0.428	0.914	0.716	0.870	0.919	0.920	0.979	0.931	0.951	0.972	0.971
crimp_st_cable_mount_box	0.18	0.442	0.372	0.745	0.291	0.363	0.485	0.589	0.713	0.931	0.749	0.933	0.811	0.969	0.659	0.961
dc_power_connector	0.661	0.726	0.661	0.725	0.849	0.507	0.995	0.770	0.720	0.921	0.715	0.950	0.922	0.947	0.944	0.994
ethernet_connector	0.967	0.853	0.981	0.866	1	0.656	1.000	0.906	0.947	0.956	0.983	0.978	0.996	0.970	0.997	0.992
ferrite_bead	0.529	0.914	0.408	0.806	0.634	0.717	0.894	0.817	0.913	0.932	0.965	0.966	0.967	0.978	0.717	0.993
fork_crimp_terminal	0.462	0.657	0.416	0.945	0.422	0.62	0.595	0.773	0.769	0.952	0.780	0.964	0.819	0.946	0.978	0.991
fuse_holder	0.853	0.861	0.564	0.957	0.309	0.605	0.597	0.754	0.736	0.927	0.770	0.948	0.866	0.915	0.998	0.996
headphone_jack_socket	0.996	0.914	0.933	0.879	0.607	0.633	0.660	0.696	0.919	0.942	0.982	0.982	0.994	0.987	0.942	0.975
humidity_sensor	0.781	0.836	0.737	0.89	0.644	0.562	0.565	0.723	0.689	0.933	0.717	0.958	0.78	0.969	0.744	0.991
knob_cap	0.637	0.893	0.672	0.879	0.656	0.425	0.919	0.656	0.903	0.958	0.925	0.938	0.931	0.947	0.923	0.976
lattice_block_plug	0.833	0.852	0.79	0.898	0.769	0.776	0.842	0.919	0.911	0.923	0.917	0.958	0.939	0.941	0.892	0.969
lego_pin_connector_plate	0.828	0.877	0.857	0.947	0.361	0.482	0.847	0.629	0.662	0.759	0.681	0.734	0.891	0.889	0.981	0.980
lego_propeller	0.615	0.739	0.939	0.799	0.348	0.62	0.471	0.703	0.540	0.727	0.530	0.773	0.739	0.863	1.000	0.933
limit_switch	0.846	0.95	0.823	0.79	0.763	0.545	0.804	0.641	0.822	0.938	0.863	0.966	0.925	0.984	0.961	0.971
miniature_lifting_motor	0.402	0.799	0.402	0.76	0.717	0.435	0.766	0.467	0.948	0.962	0.975	0.991	0.823	0.961	0.604	0.838
power_jack	0.354	0.664	0.176	0.489	0.433	0.687	0.564	0.645	0.981	0.923	0.996	0.902	0.973	0.947	0.595	0.862
purple_clay_pot	0.343	0.571	0.343	0.938	0.869	0.271	0.635	0.445	0.921	0.961	0.944	0.953	0.962	0.922	0.871	0.997
telephone_spring_switch	0.575	0.91	0.627	0.916	0.771	0.413	0.951	0.551	0.827	0.944	0.856	0.936	0.934	0.957	0.904	0.987
Avg	0.645	0.808	0.659	0.843	0.644	0.554	0.693	0.650	0.812	0.905	0.841	0.922	0.890	0.937	0.873	0.961

shows that our method yields a more separable anomaly score distribution, with a larger margin and less overlap between normal and anomalous samples. These qualitative results align with the quantitative findings, further validating **PIRN**'s effectiveness in data-scarce scenarios.

Analysis of Prototype-based Normality Encoding To better interpret how PIRN's prototypes encode normality, we added a new OT-movement visualization in Fig. 4. For several MVTec-3D-AD categories (e.g., bagel, peach) and both RGB and surface-normal branches, we visualize the displacement of patch tokens from their initial feature state (z_{pre}) to their state after BPA+APR+MNC reconstruction (z_{post}). We project prototypes and tokens into a shared 2D PCA space and draw the displacement vectors ($\Delta = z_{\text{post}} - z_{\text{pre}}$). In the plots, gray crosses denote prototypes, translucent lines show per-token movements, and bold arrows indicate the average movement of normal (green) and anomalous (red) tokens.

The visualization reveals a consistent pattern. BPA encourages prototypes to act as stable anchors for distinct normal patterns. Normal tokens start close to prototype clusters and undergo short movements during reconstruction, indicating the prototype codebook effectively approximates in-distribution patterns. In contrast, anomalous tokens lie farther away and require larger displacements toward

normal prototypes. On average, anomalous tokens exhibit 40–50% larger displacement; for example, in the RGB branch of *bagel*, $\|\Delta\|_{\text{normal}} \approx 6.2$ vs. $\|\Delta\|_{\text{anomaly}} \approx 9.0$. The accompanying displacement histograms further show that normal and anomalous images form almost non-overlapping distributions, with anomalous images consistently shifted to higher $\|\Delta\|_2$ values. This confirms that our prototype-based reconstruction induces strong normal/anomalous discrimination at the feature level.

Experiment results on RealIAD-D3 dataset. We conducte comprehensive experiments on the challenging Real-IAD D3 Zhu et al. (2025) dataset in the full-data training setting. Real-IAD D3 comprises real-world industrial components with diverse anomaly types and complex geometrie. We compared PIRN against various single-modality approaches (e.g., SimpleNet), established multimodal methods (e.g., M3DM), and D^{3}M Zhu et al. (2025).

The experimental results in Table 6 validate the effectiveness of the PIRN framework. Overall, PIRN achieves highly competitive performance, securing the best overall anomaly localization (P-AUROC) of 0.961 and the second-best overall anomaly detection (I-AUROC) of 0.873. Notably, PIRN achieves the highest P-AUROC in 13 out of the 20 categories. In terms of I-AUROC, PIRN achieves a strong score of 0.873, closely following D^{3}M (0.890). However, D^{3}M is specifically designed to leverage the unique D^3 data representation (combining 2D, Pseudo-3D, and 3D inputs). In contrast, PIRN operates using only two modalities: RGB images and derived Surface Normals (RGB + SN). Despite utilizing a simpler input representation, PIRN maintains competitive detection rates while delivering superior localization accuracy. Furthermore, PIRN substantially outperforms D^{3}M in several categories, such as 'fork_crimp_terminal' (0.978 vs. 0.819 I-AUROC) and 'lego_propeller' (1.000 vs. 0.739 I-AUROC).

Ablations. We validate each proposed module on the MVTec-3D-AD dataset, with results in Tab. 2. The baseline model (first row), similar to INP-Former Luo et al. (2025a), excludes all proposed modules. The full **PIRN** model achieves superior performance. Ablating each module from the full model results in a consistent performance drop, validating the contribution of every component. BPA contributes significantly by preventing prototype collapse, while adding APR further improves performance. The largest drop occurs when MNC is removed, highlighting the crucial role of cross-modal collaboration.

Tab. 3 evaluates prototype assignment methods in BPA. Our Balanced Optimal Transport (OT) achieves the highest performance across all metrics (e.g., AUROC$_I$ 92.2%, PRO 96.6%), outperforming alternative strategies. Softmax and linear attention yield the weakest results (AUROC$_I$ $< 85\%$), suggesting prototype under-utilization due to unconstrained assignment. Sigmoid attention performs better (AUROC$_I$ 87.8%) but still falls short.

Tab. 4 compares token aggregation strategies in APR. Among them, global averaging performs worst (AUROC$_I$ 91.5%), suggesting indiscriminate token pooling is suboptimal. Top-k averaging improves performance (AUROC$_I$ 92.1%), while our Balanced OT achieves the best results (AUROC$_I$ 92.2%, AUPRO 96.6%). The slight gain over top-k averaging indicates that balanced token contributions enable more consistent prototype refinement.

5 Conclusion

We introduced **PIRN**, a novel framework for few-shot multi-modal anomaly detection that unifies prototype-based intra-modal reconstruction with cross-modal normality communication. **PIRN** robustly models normality from scarce data via an adaptive prototype codebook. Its effectiveness comes from three key innovations: Balanced Prototype Assignment (BPA) utilizes optimal transport to mitigate codebook collapse; Adaptive Prototype Refinement (APR) dynamically adapts prototypes during inference to bridge the train-test distribution gap; and Multi-modal Normality Communication (MNC) facilitates the exchange of high-level normality cues across modalities. Extensive evaluations across MVTec 3D-AD, Eyecandies, and Real-IAD D3 demonstrate that **PIRN** establishes significant performance gains, particularly in challenging few-shot settings.

REFERENCES

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1612.00410

Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP*, pp. 202–213, 2022.

Luca Bonfiglioli, Marco Toschi, Davide Silvestri, Nicola Fioraio, and Daniele De Gregorio. The eyecandies dataset for unsupervised multimodal anomaly detection and localization. In *16th Asian Conference on Computer Vision (ACCV)*, pp. 459–475, 2022.

Yu-Min Chu, Liu Chieh, Ting-I Hsieh, Hwann-Tzong Chen, and Tyng-Luh Liu. Shape-guided dual-memory learning for 3d anomaly detection. 2023.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Alex Costanzino, Pierluigi Zama Ramirez, Giuseppe Lisanti, and Luigi Di Stefano. Multimodal industrial anomaly detection by crossmodal feature mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17234–17243, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17481–17490, 2023.

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*, pp. 1705–1714, 2019.

Hewei Guo, Liping Ren, Jingjing Fu, Yuwang Wang, Zhizheng Zhang, Cuiling Lan, Haoqian Wang, and Xinwen Hou. Template-guided hierarchical feature restoration for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6447–6458, 2023.

Liren He, Zhengkai Jiang, Jinlong Peng, Wenbing Zhu, Liang Liu, Qiangang Du, Xiaobin Hu, Mingmin Chi, Yabiao Wang, and Chengjie Wang. Learning unified reference representation for unsupervised multi-class anomaly detection. In *European Conference on Computer Vision*, pp. 216–232. Springer, 2024.

Eliahu Horwitz and Yedid Hoshen. Back to the feature: classical 3d features are (almost) all you need for 3d anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2968–2977, 2023.

Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *European Conference on Computer Vision*, pp. 303–319. Springer, 2022.

Xinlei Huang, Zhiqi Ma, Dian Meng, Yanran Liu, Shiwei Ruan, Qingqiang Sun, Xubin Zheng, and Ziyue Qiao. Praga: prototype-aware graph adaptive aggregation for spatial multi-modal omics analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 326–333, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kaifang Long, Guoyang Xie, Lianbo Ma, Jiaqi Liu, and Zhichao Lu. Revisiting multimodal fusion for 3d anomaly detection from an architectural perspective. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI*, 2025a.

Kaifang Long, Guoyang Xie, Lianbo Ma, Jiaqi Liu, and Zhichao Lu. Revisiting multimodal fusion for 3d anomaly detection from an architectural perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12273–12281, 2025b.

Ruiying Lu, YuJie Wu, Long Tian, Dongsheng Wang, Bo Chen, Xiyang Liu, and Ruimin Hu. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 36:8487–8500, 2023.

Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 9974–9983, June 2025a.

Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9974–9983, 2025b.

Kai Mao, Ping Wei, Yiyang Lian, Yangyang Wang, and Nanning Zheng. Beyond single-modal boundary: Cross-modal anomaly detection through visual prototype and harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 2025, pp. 9964–9973, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Frederik Pahde, Mihai Puscas, Tassilo Klein, and Moin Nabi. Multimodal prototypical networks for few-shot learning. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021. URL https://arxiv.org/abs/2011.08899.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. doi: 10.1561/2200000073.

Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Asymmetric student-teacher networks for industrial anomaly detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV*, pp. 2592–2602, 2023.

Sangwoo Seo, Sungwon Kim, and Chanyoung Park. Interpretable prototype-based graph information bottleneck. *Advances in Neural Information Processing Systems*, 36:76737–76748, 2023.

Pin Tang, Hai-Ming Xu, and Chao Ma. Prototransfer: Cross-modal prototype transfer for point cloud segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3337–3347, 2023.

Long Tian, Hongyi Zhao, Ruiying Lu, Rongrong Wang, Yujie Wu, Liming Wang, Xiongpeng He, and Xiyang Liu. Foct: Few-shot industrial anomaly detection with foreground-aware online conditional transport. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6241–6249, 2024.

Yuanpeng Tu, Boshen Zhang, Liang Liu, Yuxi Li, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Cairong Zhao. Self-supervised feature adaptation for 3d industrial anomaly detection. In *European Conference on Computer Vision*, pp. 75–91. Springer, 2024.

Yuanpeng Tu, Boshen Zhang, Liang Liu, Yuxi Li, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Cairong Zhao. Self-supervised feature adaptation for 3d industrial anomaly detection. In *European Conference on Computer Vision*, pp. 75–91. Springer, 2025.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1710.10903.

Fuyun Wang, Tong Zhang, Yuanzhi Wang, Yide Qiu, Xin Liu, Xu Guo, and Zhen Cui. Distribution prototype diffusion learning for open-set supervised anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20416–20426, 2025.

Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8032–8041, June 2023.

Yujie Wei, Jiaxin Ye, Zhizhong Huang, Junping Zhang, and Hongming Shan. Online prototype learning for online continual learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 18764–18774, 2023.

Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. *Advances in Neural Information Processing Systems*, 37:32111–32136, 2024a.

Yilan Zhang, Yingxue Xu, Jianqi Chen, Fengying Xie, and Hao Chen. Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. *arXiv preprint arXiv:2401.01646*, 2024b.

Chuanxia Zheng and Andrea Vedaldi. Online clustered codebook. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22798–22807, 2023.

Wenbing Zhu, Lidong Wang, Ziqing Zhou, Chengjie Wang, Yurui Pan, Ruoyi Zhang, Zhuhao Chen, Linjie Cheng, Bin-Bin Gao, Jiangning Zhang, et al. Real-iad d3: A real-world 2d/pseudo-3d/3d dataset for industrial anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15214–15223, 2025.