DUALBP-AD: A DUAL-PATH BACK-PROJECTION FRAMEWORK FOR UNIVERSAL 3D ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

3D anomaly detection serves as a crucial task in industrial applications, dedicated to identifying the anomalies and localizing the anomalous regions. However, existing methods are typically tailored to specific task configurations and particular modality inputs, thereby lacking a universal paradigm. To overcome these bottlenecks, we propose DualBP-AD, a universal dual-path back-projection framework that conceptualizes anomaly detection through a cross-modal rendering paradigm. By projecting unstructured 3D data into 2D images, our framework leverages the robust generalization capability of 2D foundational models , ensuring these extracted features maintain strong semantic power when back-projected onto 3D spatial points. To resolve contextual confusion across varying viewpoints and categories, we design an instance-adaptive spatial-semantic transition layer. Furthermore, we introduce an orthogonal dual-path back-projection mechanism to accurately map and fuse multi-view anomaly back into the native 3D topology. Extensive experiments demonstrate that DualBP-AD overcomes task-specific and data-specific limitations, yielding highly accurate detection and precise localization across single-class, multi-class, few-shot, and zero-shot settings on four benchmarks of diverse modalities. For example, it achieves 95.6% O-AUROC and 97.3% P-AUROC on Anomaly-ShapeNet, outperforming the previous SOTA.

1 INTRODUCTION

3D anomaly detection is critical for identifying structural deformations and surface defects in industrial applications. Unlike traditional 2D images, 3D data provides comprehensive geometric context, making it indispensable for practical industrial applications. However, establishing a universal detection framework across diverse object geometries and data modalities remains a challenge.

The development of such a universal framework is constrained by native 3D pre-trained models. Due to the severe scarcity of large-scale 3D datasets, existing 3D backbones learn rigid structural priors primarily from synthetic, noise-free data (Koch et al., 2019; Wu et al., 2021; Dong et al., 2026). Consequently, they lack the representational capacity to classify normal geometric variations from actual physical defects (Gu et al., 2023; Cabrera et al., 2024). This leads to "over-smoothing" and the "identity mapping" of highly localized anomalies during feature reconstruction. Furthermore, current approaches are typically constrained by restricted task configurations. They rely on strict task settings or modality inputs, which limits their applicability in complex and dynamic industrial environments. When operational conditions shift, these specialized methods fail to generalize. This vulnerability highlights the crucial need for a unified detection paradigm.

To overcome these fundamental bottlenecks, we introduce DualBP-AD, a universal dual-path back-projection framework. Rather than relying on biased 3D backbones, we conceptualize the anomaly detection problem through a cross-modal rendering paradigm. We project unstructured 3D spatial data into dense 2D images, directly exploiting the robust generalization capability of 2D foundational models. To prevent contextual confusion across varying viewpoints and categories, we establish an instance-adaptive spatial-semantic transition layer. This layer utilizes feature recentering and dynamic query generation to accurately align multi-view 2D features with specific instance semantics. Crucially, we introduce tailored 3D modules, including Gaussian spatial aggregation and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

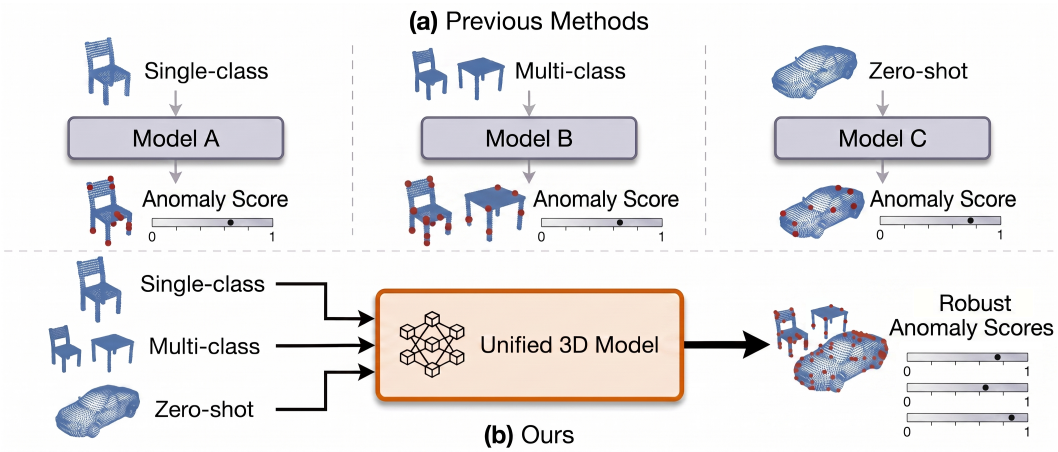


Figure 1: A Unified Framework for 3D Anomaly Detection.

curvature-aware 3D soft mining. These mechanisms explicitly preserve physical continuity and mitigate multi-view alignment errors before aggregating detection cues back into the 3D space.

The main contributions of this work are summarized as follows:

- We propose a dual-path back-projection mechanism coupled with a two-stage optimization strategy. The 2D and 3D detection paths precisely capture anomaly representations at both the pixel and point levels. Additionally, the pre-training and post-training optimization paradigm ensures robust adaptability across diverse input modalities.
- We introduce a universal framework for 3D anomaly detection. To the best of our knowledge, it is the first unified approach designed to be readily applicable across various task configurations (single-class, multi-class, few-shot, and zero-shot) and diverse input modalities (point clouds, XYZ coordinates, depths and surface normals).
- We conduct extensive experiments on multiple benchmarks under diverse task configurations. The results demonstrate that our approach achieves SOTA performance in the majority of experiments for both object-level and point-level.

2 Related Work

2.1 2D Anomaly Detection

Anomaly detection in the 2D domain has substantial progress, propelled by methodological innovations alongside the robust representational capabilities of pre-trained backbone networks (Yang et al., 2025; Caron et al., 2021; Oquab et al., 2024; Kirillov et al., 2023). Existing algorithms follow three dominant paradigms. The memory bank paradigm identifies anomalies by calculating feature distances to stored normal embeddings (Roth et al., 2022; Defard et al., 2020; Xu et al., 2026). The pseudo-anomaly regression paradigm synthesizes local defects to train discriminative networks (Li et al., 2021; Liu et al., 2023b). Finally, the reconstruction-based paradigm utilizes autoencoders or generative models to reconstruct normal patterns, leveraging reconstruction discrepancies as anomaly scores (Gong et al., 2019; He et al., 2023).

2.2 3D Anomaly Detection

3D anomaly detection has emerged to capture complex spatial geometries. Initial approaches utilized native 3D backbones to extract features from raw point clouds (Zhao et al., 2021; Wu et al., 2022; 2024). However, being pre-trained on idealized synthetic datasets, these models learn structural priors that conflict with localized physical defects, inevitably causing severe over-smoothing. To overcome these limitations, recent works typically pursue two distinct paths: some methods

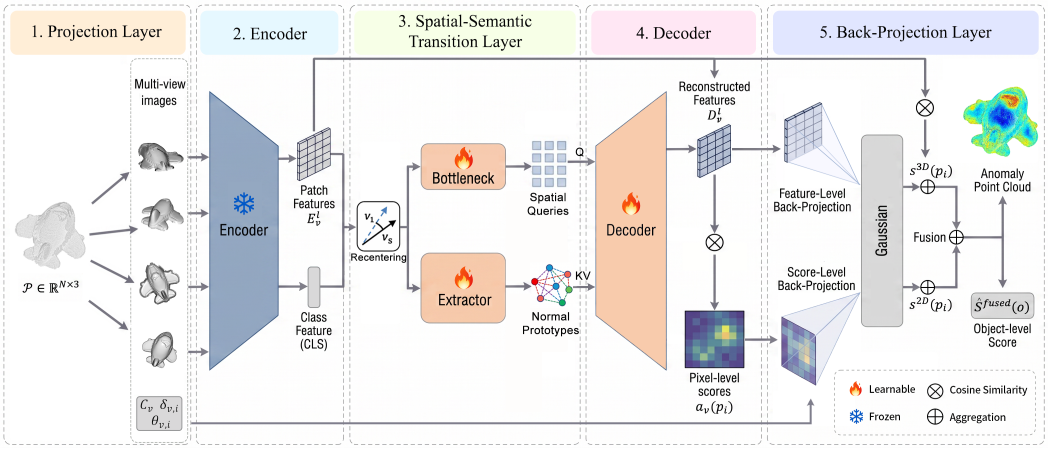


Figure 2: Overall architecture of DualBP-AD. The framework features a dual-path back-projection mechanism and is optimized via a two-stage training strategy for universal 3D anomaly detection.

completely discard pre-trained 3D models for end-to-end training paradigms Chen et al. (2023); Zavrtanik et al. (2024), whereas others project 3D spatial data into 2D representations.

2.3 Unified Models

In contemporary anomaly detection, there is a growing transition from task-specific architectures to unified models. Recent advancements in the 2D domain have successfully established unified frameworks capable of evaluating multiple categories simultaneously (Guo et al., 2025b;a; Gu et al., 2025). However, the 3D anomaly detection domain remains largely blank in this regard, lacking a universal paradigm effective across single-class, multi-class, few-shot, and zero-shot environments. Advanced unified configurations require robust generalization to unseen objects with minimal or no reference data. Consequently, leveraging cross-modal representations and instance-adaptive mechanisms emerges as a critical strategy to achieve universality in 3D detection tasks.

3 Method

3.1 Overall Architecture

The primary objective of 3D anomaly detection is to localize structural and surface anomalies within an input 3D data $\mathcal{P}$ comprising N spatial coordinates p_i. The overall forward pass progresses sequentially through five core modules: the Projection Layer, the Encoder, the Spatial-Semantic Transition Layer, the Decoder, and the Back-Projection Layer. For notational brevity, the following methodology is formalized using point cloud representations. The explicit projection and back-projection operations are not required when processing structured 3D data, yet the core anomaly detection mechanism remains completely identical.

The framework's data flow proceeds as follows: the Projection Layer renders the raw input $\mathcal{P}$ from V distinct viewpoints into a set of discrete 2D views I_v while establishing deterministic pixel-to-point correspondence maps $\mathcal{C}_v$. A pre-trained Encoder then processes I_v to extract multi-level semantic features $\mathbf{E}_v^l$. To bridge the latent space, the Spatial-Semantic Transition Layer transforms these features to formulate spatial queries ($\mathbf{Q}$) and dynamically extracts a set of Intrinsic Normal Prototypes, providing the keys and values ($\mathbf{K}, \mathbf{V}$). Guided by these prototypes, the Decoder reconstructs normal feature representations $\mathbf{D}_v^l$. The reconstruction discrepancy inherently isolates anomalous regions, yielding per-view anomaly scores $a_v(p_i)$. Crucially, the Back-Projection Layer utilizes $\mathcal{C}_v$ to map multi-view outputs back into the 3D coordinate space. This reverse mapping aggregates the per-view scores into a geometrically robust 2D anomaly score $s^{2\mathrm{D}}(p_i)$. Concurrently, it fuses the multi-view features to form the 3D point-level encoder and decoder features, denoted as $\tilde{\mathbf{e}}^l(p_i)$ and $\tilde{\mathbf{d}}^l(p_i)$, the discrepancy of which yields the 3D anomaly score $s^{3\mathrm{D}}(p_i)$. The point-level scores are

subsequently fused to produce the final object-level prediction $\hat{S}^{\text{fused}}(o)$. The entire architecture is optimized via a structured two-stage training strategy that balances 2D semantic reconstruction and 3D spatial alignment.

3.2 PROJECTION LAYER

While 2D foundational models like DINOv3 (Siméoni et al., 2025) and SAM2 (Ravi et al., 2024) exhibit exceptional zero-shot generalization, the 3D domain lacks a backbone network with comparable universality. Recent 3D pre-training models, such as PointMAE (Pang et al., 2022), show fundamental limitations in anomaly detection. Their structural priors—learned from idealized, noise-free synthetic datasets—inherently conflict with the highly localized geometric deviations of real-world anomalies. Applying these models often induces over-smoothing and the identity mapping of anomalies. To circumvent these biases and fully exploit the representational power of 2D models, our framework adopts a multi-view projection and back-projection paradigm.

As the interface of our architecture, the Projection Layer systematically renders the input 3D point cloud $\mathcal{P}$ into 2D representations. By centralizing the object within a global coordinate system, V virtual cameras, each governed by an extrinsic matrix $\mathbf{E}_v^{\text{cam}}$, capture distinct viewpoints $v \in \{1, \ldots, V\}$. This process yields a set of multi-view RGB images I_v, effectively transforming unstructured 3D spatial topologies into dense 2D grids optimized for feature extraction.

Beyond visual rendering, this layer establishes deterministic geometric mappings critical for subsequent 3D back-projection. For each view v, a pixel-to-point correspondence map $\mathcal{C}_v$ records the projected 2D coordinates $(u_{v,i}, w_{v,i})$ of any given 3D point p_i. A binary visibility indicator $\delta_{v,i}$ defines the valid visible view set $\mathcal{V}(i) = \{v \mid \delta_{v,i} = 1\}$ for each point. We calculate the incidence angle $\theta_{v,i}$ between the point's surface normal $\mathbf{n}_i$ and the camera viewing ray. The geometric metadata is essential for resolving visual ambiguities and aggregating 2D information back into the original 3D topology.

3.3 SPATIAL-SEMANTIC TRANSITION LAYER

To extract representations from the rendered views I_v, we employ a pre-trained and frozen 2D foundation model (e.g., DINOv3) as the Feature Encoder, generating multi-level semantic patch tokens $\mathbf{E}_v^l$. The Spatial-Semantic Transition Layer operates exclusively within the latent space, serving as the foundational bridge connecting the Feature Encoder and the Decoder. This layer splits into two parallel branches: a Bottleneck module dedicated to spatial transformations and a Prototype Extraction module dedicated to semantic aggregation.

A primary challenge in cross-category anomaly detection is context confusion, where identical local features may represent normal structures in one category but anomalies in another. This problem is further exacerbated by the significant appearance discrepancies across different multi-views of the same instance. To mitigate the ambiguity inherent in a unified feature space, we adapt a Class Token (CLS) Feature Recentering mechanism. We leverage the ViT Class Token (Dosovitskiy et al., 2021), which encodes the category-specific and view-specific contextual patterns, as a contextual anchor. Before routing patch tokens $\mathbf{f}_n$ to the branches, they are projected into a localized reference frame where their respective Class Token serves as the new zero point. Through this contextual recentering, the feature space accurately aligns local geometric topologies with global semantic context.

The spatial branch acts as an information bottleneck. It applies a standard MLP to compress the channel dimensions of the recentered patch features. This straightforward bottleneck abstracts the spatial hierarchies and filters redundant background noise, efficiently formulating the spatially aligned features required for subsequent feature reconstruction.

The semantic branch dynamically extracts the Intrinsic Normal Prototypes, denoted as the matrix $\mathbf{K} \in \mathbb{R}^{M \times C}$ comprising prototype vectors $\mathbf{k}_m$. Inspired by INP-Former (Luo et al., 2025a;b), we optimize its static query mechanism. The previous mechanism has a critical limitation: because the queries are globally shared and fixed during inference, they struggle to adapt to the substantial geometric and scale variances across different categories, instances, and viewpoints. To address this, we propose an instance-adaptive dynamic query. Specifically, our queries combine the original static learnable tokens with a dynamic bias derived from the image-specific Class Token. By injecting this context-dependent bias, the queries are uniquely tailored to each input image. They then attend to

the recentered patch tokens $\mathbf{f}_n$ to aggregate representative normal patterns. This process formulates the INP matrix $\mathbf{K}$, accurately capturing the normal patterns specific to the evaluated instance.

Following the latent transition, the Decoder receives the spatially aligned features from the Bottleneck branch (serving as query inputs) and the Intrinsic Normal Prototypes from the Prototype Extraction branch (serving as key and value inputs) to perform cross-attention. The feature reconstruction process is explicitly guided by these normal prototypes. Because the model strictly learns to extract normal prototypes and reconstruct normal representations during training, this design ensures that normal regions are accurately reconstructed during inference, while anomalous regions inherently fail to be reconstructed. Consequently, this mechanism effectively resolves the identity mapping problem. Ultimately, the Decoder layer yields two pixel-level components for each view: the reconstructed features $\mathbf{D}_v^l$ and the pixel-wise anomaly scores $a_v(p_i)$, which are directly derived from the reconstruction discrepancy.

3.4 Back-Projection Layer

As the inverse of rendering, this layer aggregates multi-view results back into the 3D space. Utilizing the correspondence map $\mathcal{C}_v$, valid spatial properties from the visible view set $\mathcal{V}(i)$ are mapped to each 3D point $p_i \in \mathcal{P}$. This reverse mapping employs a dual-path mechanism: score-level and feature-level alignments.

During inference, the Score-Level Back-Projection formulates a robust 2D point-level anomaly score $s^{2\mathrm{D}}(p_i)$. To mitigate perspective distortions, we introduce a geometric reliability weight $W^{\mathrm{geom}}(p_i, v) = \max\big(|\cos\theta_{v,i}|, \delta_{\mathrm{floor}}\big)^{\gamma}$, prioritizing views orthogonal to the surface normal $\mathbf{n}_i$. Selecting the top-s highest-scoring visible views $\mathcal{T}_s(i)$, the pixel-wise anomaly score $a_v(p_i)$ and its back-projected counterpart $s^{2\mathrm{D}}(p_i)$ are defined in Eq. equation 1 and Eq. equation 2. This mechanism effectively filters out projection artifacts.

$$a_v(p_i) = \frac{1}{L}\sum_{l}\big(1 - \cos(\mathbf{e}_v^l(p_i), \mathbf{d}_v^l(p_i))\big) \tag{1}$$

$$s^{2\mathrm{D}}(p_i) = \frac{\sum_{v\in\mathcal{T}_s(i)} W^{\mathrm{geom}}(p_i, v)\cdot a_v(p_i)}{\sum_{v\in\mathcal{T}_s(i)} W^{\mathrm{geom}}(p_i, v)} \tag{2}$$

The Feature-Level Back-Projection maps latent representations back to 3D space during both training and inference. For each feature channel c at point p_i, representations across visible views $\mathcal{V}(i)$ are aggregated via a dynamic pooling function $\mathrm{pool}(\cdot)$. Applying this to the generic multi-view features $\mathbf{h}_{v,c}^l(p_i)$ yields the fused 3D encoder and decoder features ($\tilde{\mathbf{e}}^l(p_i)$ and $\tilde{\mathbf{d}}^l(p_i)$). The 3D anomaly score $s^{3\mathrm{D}}(p_i)$ is then derived via their mean cosine distance across L levels, as formalized in Eq. equation 3. $\mathrm{pool}(\cdot)$ is instantiated as max-pooling during inference to explicitly highlight structural deviations, and as mean-pooling during training to ensure stable gradient updates.

$$\tilde{\mathbf{h}}^l(p_i)_c = \mathrm{pool}_{v\in\mathcal{V}(i)}\big(\mathbf{h}_{v,c}^l(p_i)\big), \quad s^{3\mathrm{D}}(p_i) = \frac{1}{L}\sum_{l}\Big(1 - \cos\Big(\tilde{\mathbf{e}}^l(p_i), \tilde{\mathbf{d}}^l(p_i)\Big)\Big) \tag{3}$$

To reflect the physical continuity of real-world anomalies and suppress isolated noise, we introduce Gaussian Spatial Aggregation (G-aggregation). For each point p_i, a Gaussian weight $w_{i,j}^{\mathrm{gauss}}$ is assigned to its K-nearest neighbors $\mathcal{N}_K(p_i)$. Let $\mathcal{X}(p_i) \in \{s^{2\mathrm{D}}(p_i), s^{3\mathrm{D}}(p_i), \tilde{\mathbf{e}}^l(p_i)\}$ denote the target property, which is refined into its smoothed counterpart $\hat{\mathcal{X}}(p_i)$ according to Eq. equation 5:

$$w_{i,j}^{\mathrm{gauss}} = \exp\left(-\frac{||p_i - p_j||_2^2}{2\sigma^2}\right) \tag{4}$$

$$\hat{\mathcal{X}}(p_i) = 0.5\cdot\mathcal{X}(p_i) + 0.5\cdot\frac{\sum_{p_j\in\mathcal{N}_K(p_i)} w_{i,j}^{\mathrm{gauss}}\cdot\mathcal{X}(p_j)}{\sum_{p_j\in\mathcal{N}_K(p_i)} w_{i,j}^{\mathrm{gauss}}} \tag{5}$$

Finally, the smoothed point-level scores are aggregated into branch-specific object-level scores, $\hat{S}^{2\mathrm{D}}(o)$ and $\hat{S}^{3\mathrm{D}}(o)$, by averaging their respective top-$z\%$ values. The fused object-level score $\hat{S}^{\mathrm{fused}}(o)$ is computed via a fusion weight α, as defined in Eq. equation 6:

$$\hat{S}^{\mathrm{fused}}(o) = \alpha\cdot\hat{S}^{3\mathrm{D}}(o) + (1-\alpha)\cdot\hat{S}^{2\mathrm{D}}(o) \tag{6}$$

3.5 LOSS FUNCTION

To extract normal prototypes, we employ the Soft INP Coherence Loss ($\mathcal{L}_{\text{sc}}$). This loss reconstructs each patch token $\mathbf{f}_n$ as a cosine-similarity-weighted mixture of all INPs $\mathbf{k}_m$. Minimizing the resulting error ensures the extracted INPs faithfully represent normal patterns by pulling normal tokens toward the INP manifold:

$$\hat{\mathbf{f}}_n = \sum_m \text{softmax}_m\big(\cos(\mathbf{f}_n, \mathbf{k}_m)\big)\mathbf{k}_m, \quad \mathcal{L}_{\text{sc}} = \frac{1}{N_p}\sum_n \big(1 - \cos(\mathbf{f}_n, \hat{\mathbf{f}}_n)\big) \tag{7}$$

To guide the Decoder in accurately reconstructing normal patterns across multiple semantic scales, we formulate dual-level reconstruction losses. For 2D semantic reconstruction, the Pixel-Level Reconstruction Loss ($\mathcal{L}^{\text{2D}}_{\text{recon}}$) computes the average cosine distance between the original encoder token $\mathbf{f}^l_{v,n}$ and its reconstructed counterpart $\mathbf{d}^l_{v,n}$. Crucially, this training loss is exactly the pixel-wise anomaly score $a_v(p_i)$ used during inference, as defined in Eq. equation 1. To enforce 3D topological consistency, the Point-Level 3D Reconstruction Loss ($\mathcal{L}^{\text{3D}}_{\text{recon}}$) minimizes the distance between the 3D-fused encoder and decoder features ($\tilde{\mathbf{e}}^l(p_i)$ and $\tilde{\mathbf{d}}^l(p_i)$). It shares the same formulation as $s^{\text{3D}}(p_i)$ in Eq. equation 3, with the key distinction that the training loss utilizes mean-pooling to aggregate, whereas the inference score employs channel-wise max-pooling. The two reconstruction losses are defined in Eq. equation 8 and Eq. equation 9:

$$\mathcal{L}^{\text{2D}}_{\text{recon}}(v, n) = \frac{1}{L}\sum_l \big(1 - \cos(\mathbf{f}^l_{v,n}, \mathbf{d}^l_{v,n})\big) \tag{8}$$

$$\mathcal{L}^{\text{3D}}_{\text{recon}}(p_i) = \frac{1}{L}\sum_l \Big(1 - \cos\Big(\tilde{\mathbf{e}}^l(p_i), \tilde{\mathbf{d}}^l(p_i)\Big)\Big) \tag{9}$$

To address varying optimization difficulties and prevent false positives, we apply a difficulty-aware Soft Mining mechanism to dynamically modulate gradients across both modalities. During training on exclusively nominal data, the two branches encounter distinct optimization bottlenecks. Specifically, the 2D soft mining targets complex appearances and intricate semantic textures that naturally yield higher reconstruction errors, while the 3D soft mining focuses on difficult-to-fit geometric topologies and structural variations. To counter this, we introduce a unified dynamic weight defined as the ratio of the local reconstruction error to the mean error (μ_{2D} or μ_{3D}), parameterized by a focusing exponent (γ_{2D} or γ_{3D}). Crucially, this weight is strictly treated as a non-differentiable constant (stop-gradient) during backpropagation. This mechanism forces the model to strictly penalize underfitting on hard-to-reconstruct normal patterns in both branches without disrupting the global feature manifold (Further details are provided in the appendix). The final modulated training losses for the 2D and 3D branches are aggregated across all views and visible points, as formulated in Eq. equation 10 and Eq. equation 11, respectively:

$$\mathcal{L}_{\text{2D}} = \frac{1}{V \cdot N_p}\sum_{v,n}\left(\frac{\mathcal{L}^{\text{2D}}_{\text{recon}}(v, n)}{\mu_{\text{2D}}}\right)^{\gamma_{\text{2D}}} \cdot \mathcal{L}^{\text{2D}}_{\text{recon}}(v, n) \tag{10}$$

$$\mathcal{L}_{\text{3D}} = \frac{1}{|\tilde{\mathcal{P}}_{\text{vis}}|}\sum_{p_i}\left(\frac{\mathcal{L}^{\text{3D}}_{\text{recon}}(p_i)}{\mu_{\text{3D}}}\right)^{\gamma_{\text{3D}}} \cdot \mathcal{L}^{\text{3D}}_{\text{recon}}(p_i) \tag{11}$$

To stabilize the optimization process, the framework employs a two-stage training strategy. During the pre-training stage (E_1 epochs), the model focuses on learning basic 2D semantic reconstructions and extracting prototypes, with the prototype coherence loss ($\mathcal{L}_{\text{sc}}$) modulated by a weighting parameter λ_2. In the subsequent post-training stage (E_2 epochs), the network is fine-tuned to align the 2D representations with 3D topologies by integrating the 3D auxiliary loss, governed by a fusion weight λ_1. The optimization objectives for both stages are defined in Eq. equation 12 and Eq. equation 13:

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{2D}} + \lambda_2\mathcal{L}_{\text{sc}} \tag{12}$$

$$\mathcal{L}_{\text{post}} = \mathcal{L}_{\text{2D}} + \lambda_1\mathcal{L}_{\text{3D}} + \lambda_2\mathcal{L}_{\text{sc}} \tag{13}$$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. The efficacy of the proposed framework is evaluated on four benchmark datasets: Anomaly-ShapeNet (Li et al., 2023), Real3D-AD (Liu et al., 2023a), MVTec 3D-AD (Bergmann et al., 2022), and Eyecandies (Bonfiglioli et al., 2022). Anomaly-ShapeNet contains 1,472 point cloud samples across 40 object categories with 7 distinct anomaly types. Real3D-AD comprises 1,254 samples across 12 categories, featuring 2 anomaly types. MVTec 3D-AD is a multi-modal dataset encompassing 4,147 samples across 10 categories with 9 anomaly variants. Eyecandies contains 15,000 samples across 10 categories with 4 anomaly variants. Since all training sets consist exclusively of normal samples while test sets encompass both normal and anomalous instances, our experiments strictly adhere to an unsupervised anomaly detection paradigm.

Evaluation Metrics. The framework is assessed across two dimensions: object-level and point-level. We report three evaluation metrics: the Area Under the Receiver Operating Characteristic Curve (AUROC), Average Precision (AP), and the maximum F1-score (F1). Among these, AUROC serves as the primary metric for evaluating overall detection performance.

Baseline Methods. We compare our framework against several recent state-of-the-art baselines, including single-class methods(Ye et al. (2024); Liang et al. (2025); Zha et al. (2025); Kang et al. (2026); Cheng et al. (2025b); Zheng et al. (2025); Horwitz & Hoshen (2023); Rudolph et al. (2023); Wang et al. (2025a); Liang et al. (2025); Cao et al. (2024a); Tao et al. (2025)), multi-class methods (Kim et al. (2026); Cheng et al. (2025a)), and zero-shot methods(Li et al. (2026); Zhou et al. (2024; 2025); Wang et al. (2025b); Cheng et al. (2026); Deng et al. (2026); Cao et al. (2024b)). The results of these methods are strictly obtained from publicly available code or their original papers.

Implementation Details. The projection layer renders images at a default resolution of 512×512. We employ DINOv3-B as the pre-trained feature encoder, extracting intermediate representations from transformer blocks $\{6, 7, 8, 9\}$. The number of Intrinsic Normal Prototypes (M) is set to 6. The optimization follows a two-stage training strategy with the epochs configured as $E_1 = 100$ and $E_2 = 50$. The loss weighting hyperparameters are established as $\lambda_1 = 0.2$ and $\lambda_2 = 0.2$. For the Gaussian spatial aggregation module, the neighborhood size is set to $K = 25$ with a smoothing bandwidth of $\sigma = 1.0$. Further details regarding the experimental setup are provided in the appendix.

Table 1: Few-shot(3-shot)/Single-class/Multi-class/Multi-class(with cls-label) detection results of O-AUROC (P-AUROC) on Real3D-AD.

Method	Airplane	Car	Candybar	Chicken	Diamond	Duck	Fish	Gemstone	Seahorse	Shell	Starfish	Toffees	Mean
Single-class Method													
Reg3D-AD(NeurIPS'23)	71.6(63.1)	69.7(71.8)	68.5(72.4)	85.2(67.6)	90.0(83.5)	58.4(50.3)	91.5(82.6)	41.7(54.5)	76.2(81.7)	58.3(81.1)	50.6(61.7)	82.7(75.9)	**70.4(70.5)**
PASDF(ICCV'25)	62.8(77.7)	95.9(80.2)	78.8(54.6)	73.9(76.8)	89.4(69.9)	65.8(78.2)	98.9(83.7)	63.4(65.4)	100.0(88.7)	85.0(64.8)	61.7(70.3)	86.6(83.8)	**80.2(74.5)**
Simple3D(AAAI'26)	76.5(88.1)	98.1(99.2)	65.1(96.2)	82.6(86.1)	100.0(99.0)	77.8(96.6)	91.2(96.2)	70.4(97.3)	93.0(94.2)	51.4(71.6)	69.6(85.8)	88.8(96.8)	**80.4(92.3)**
CASL(AAAI'26)	80.8(84.2)	79.9(90.5)	84.8(93.2)	65.7(71.3)	97.6(98.8)	83.6(89.5)	93.5(93.5)	76.9(91.6)	64.3(81.4)	79.1(87.3)	89.3(83.9)	92.4(93.7)	**82.3(88.2)**
Point-Patch Fusion(CVPR'26)	87.5(-)	75.6(-)	86.8(-)	82.3(-)	90.1(-)	91.6(-)	92.4(-)	70.1(-)	86.3(-)	83.7(-)	78.5(-)	85.9(-)	**84.2(-)**
Ours(3-shot)	66.0(94.6)	82.6(95.3)	96.9(97.4)	74.7(94.2)	99.9(98.5)	71.4(91.8)	100.0(98.0)	86.4(97.6)	95.8(97.1)	67.7(96.8)	96.7(96.5)	96.2(97.9)	**86.2(96.3)**
Ours	70.6(95.0)	85.2(95.4)	97.2(97.7)	70.3(94.0)	99.9(98.7)	67.8(92.5)	100.0(98.1)	87.2(97.7)	97.9(97.4)	90.8(97.4)	97.6(96.6)	95.2(97.5)	**88.3(96.5)**
Multi-class Method													
MC3D-AD(IJCAI'25)	85.0(62.8)	74.9(81.9)	83.0(91.0)	71.5(64.0)	95.5(94.2)	83.1(82.2)	86.5(93.2)	56.0(45.8)	71.6(65.9)	80.3(77.8)	76.6(69.0)	73.8(93.4)	**78.2(76.8)**
SeDiR(CVPR'26)	86.0(66.6)	78.3(90.8)	81.9(95.1)	72.9(63.2)	94.8(97.7)	86.2(83.4)	93.8(96.7)	62.7(54.7)	67.4(73.9)	77.9(79.8)	85.4(70.9)	84.5(94.0)	**81.0(80.6)**
Ours	61.9(95.3)	85.8(95.8)	99.4(97.6)	70.5(94.1)	100.0(98.3)	84.3(94.8)	99.3(97.8)	83.8(97.8)	97.3(97.1)	93.0(97.3)	88.1(96.8)	99.4(98.9)	**88.6(96.8)**
Ours(cls-label)	71.0(95.6)	90.4(95.6)	99.2(97.6)	67.8(94.0)	100.0(98.7)	78.8(94.9)	99.9(97.9)	85.6(98.0)	98.6(97.2)	92.3(97.4)	92.8(96.9)	98.0(98.8)	**89.5(96.9)**

4.2 CROSS-MODALITY RESULTS

As the primary evaluation, we conduct single-class experiments, training an independent model for each category. At the object level, DualBP-AD demonstrates strong detection accuracy on each evaluated dataset. The method achieves an O-AUROC of 95.6% on Anomaly-ShapeNet(previous SOTA 90.9%; see Table 2) and 88.3% on Real3D-AD(previous SOTA 84.2%; see Table 1). This performance extends to MVTec 3D-AD(97.4% compared to previous SOTA 95.4%; as shown in 3), and Eyecandies(95.3% compared to previous SOTA 90.2%; as shown in 4). Furthermore, we achieve SOTA in other metrics(96.9% AP 94.0% F1 on Anomaly-ShapeNet and 88.2% AP and 86.5% F1 on Real3D-AD). These metrics confirm robust classification efficacy across disparate modalities.

At the point level, DualBP-AD similarly achieves precise localization. The framework attains a Point-AUROC (P-AUROC) of 97.3% on Anomaly-ShapeNet (previous SOTA 92.9%), 96.5% on Real3D-AD (previous SOTA 92.3%), 99.0% on MVTec 3D-AD (previous SOTA 99.2%), and 98.9%

Table 2: Few-shot/Single-class/Multi-class/Multi-class(with cls-label) results of O-AUROC on Anomaly-ShapeNet.

Method	ashtray0	bag0	bottle0	bottle1	bottle3	bowl0	bowl1	bowl2	bowl3	bowl4	bowl5	bucket0	bucket1	cap0	cap3	cap4	cap5	cup0	cup1	eraser0	Mean
Single-class Method																					
PO3AD(CVPR'25)	100.0	83.3	90.0	93.3	92.6	92.2	82.9	83.3	88.1	98.1	84.9	85.3	78.7	87.7	85.9	79.2	67.0	87.1	83.3	99.5	**83.9**
Simple3D(AAAI'26)	99.5	88.1	97.6	95.1	100.0	100.0	83.0	71.1	91.1	73.0	86.3	95.9	79.0	85.2	86.7	91.2	80.4	100.0	82.4	100.0	**86.0**
CASL(AAAI'26)	94.3	94.8	95.7	95.4	100.0	100.0	93.3	99.3	99.6	94.4	88.8	99.0	91.4	94.8	82.1	77.2	55.4	99.0	64.8	99.5	**88.7**
PASDF(ICCV'25)	100.0	99.5	100.0	100.0	100.0	100.0	94.8	100.0	100.0	93.3	91.2	96.8	77.5	85.2	64.9	64.6	85.3	97.1	85.7	95.2	**90.7**
MC4AD(CVPR'25)	100.0	97.6	95.7	100.0	95.6	98.5	90.4	91.9	93.7	99.6	91.9	94.3	90.2	87.8	91.6	83.5	81.8	99.0	93.8	100.0	**90.9**
Ours(2-shot)	100.0	80.5	98.1	98.6	93.3	98.2	98.5	97.8	97.0	98.9	100.0	95.6	94.6	96.3	99.3	97.2	98.6	83.8	91.0	100.0	**92.5**
Ours	100.0	83.3	99.1	99.3	99.1	97.8	98.9	99.3	98.5	99.3	100.0	96.5	94.6	94.1	99.7	96.5	100.0	100.0	98.1	100.0	**95.6**
Multi-class Method																					
MC3D-AD(IJCAI'25)	96.2	80.5	79.5	70.9	75.6	93.0	97.8	71.9	88.5	91.1	75.4	89.8	78.4	79.3	70.1	83.5	76.1	74.3	95.2	77.6	**84.2**
SeDiR(CVPR'26)	97.6	88.6	92.9	91.6	98.4	95.9	95.6	92.6	96.7	98.5	97.9	91.1	93.3	98.5	98.9	99.6	87.4	99.5	100.0	66.7	**93.3**
Ours	100.0	68.6	99.5	99.3	100.0	97.8	100.0	100.0	98.2	100.0	100.0	98.7	93.3	96.7	99.3	93.3	98.3	97.1	93.8	100.0	**94.4**
Ours(cls-label)	100.0	69.5	99.5	99.3	100.0	99.3	100.0	100.0	99.3	100.0	100.0	97.1	94.0	96.7	99.7	95.1	99.7	95.2	94.8	100.0	**95.0**

Method	headset0	headset1	helmet0	helmet1	helmet2	helmet3	jar	phone	shelf0	tap0	tap1	vase0	vase1	vase2	vase3	vase4	vase5	vase7	vase8	vase9	Mean
Single-class Method																					
PO3AD(CVPR'25)	80.8	92.3	76.2	96.1	86.9	75.4	86.6	77.6	57.3	74.5	68.1	85.8	74.2	95.2	82.1	67.5	85.2	96.6	73.9	83.0	**83.9**
Simple3D(AAAI'26)	98.2	95.7	69.0	71.9	75.4	65.5	90.5	100.0	74.8	70.3	59.6	95.4	82.4	87.1	81.2	86.4	96.2	89.5	85.5	81.5	**86.0**
CASL(AAAI'26)	83.6	96.7	78.3	89.0	88.1	86.1	99.5	98.1	83.2	76.1	56.3	83.8	89.0	98.1	93.0	82.1	71.4	100.0	93.3	81.8	**88.7**
PASDF(ICCV'25)	100.0	79.5	81.2	93.8	76.5	84.6	100.0	100.0	71.3	88.2	79.3	100.0	92.9	100.0	80.6	91.2	100.0	100.0	92.4	83.6	**90.7**
MC4AD(CVPR'25)	88.4	99.0	84.1	98.6	96.2	80.3	91.4	92.4	69.6	87.3	70.0	98.3	83.8	97.6	84.8	75.5	91.0	96.7	81.8	90.9	**90.9**
Ours(2-shot)	91.1	98.1	86.7	83.3	93.6	81.8	99.5	99.1	83.5	84.2	71.9	99.6	80.5	92.9	84.9	81.2	89.5	96.2	94.9	90.9	**92.5**
Ours	98.7	100.0	93.3	91.0	92.8	89.7	100.0	100.0	88.4	90.6	77.4	100.0	93.8	100.0	90.0	84.9	95.2	97.1	97.9	90.6	**95.6**
Multi-class Method																					
MC3D-AD(IJCAI'25)	86.2	88.6	67.2	100.0	60.9	97.9	97.1	91.9	84.1	94.5	97.0	82.1	85.7	92.9	76.1	87.6	97.6	93.8	67.0	73.6	**84.2**
SeDiR(CVPR'26)	84.9	89.0	96.5	100.0	87.0	100.0	100.0	90.5	88.7	92.7	89.3	88.8	97.1	97.6	78.8	92.4	94.8	100.0	87.3	94.2	**93.3**
Ours	93.3	95.7	80.9	68.6	97.7	90.0	95.2	100.0	91.3	93.0	85.2	100.0	90.5	99.5	91.2	82.4	97.1	99.1	99.7	89.7	**94.4**
Ours(cls-label)	95.6	96.2	84.4	77.1	96.8	91.8	96.2	100.0	89.6	90.6	84.8	100.0	91.0	100.0	94.9	84.2	97.1	99.5	99.7	91.8	**95.0**

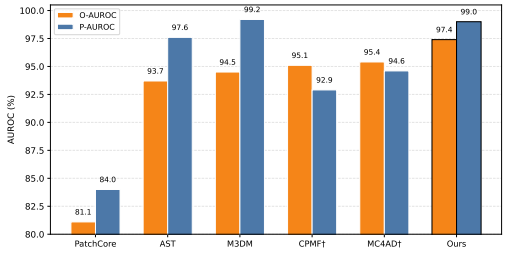


Figure 3: Single-class results on MVTec3D-AD. Methods without RGB marked with †.

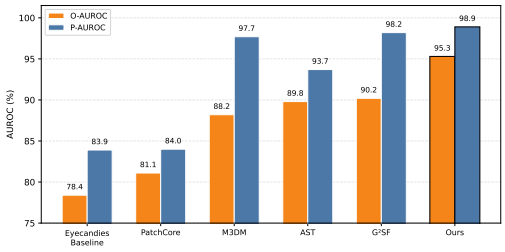


Figure 4: Single-class results on Eyecandies. We report O-AUROC and P-AUROC (%).

on Eyecandies (previous SOTA 98.2%). Notably, previous SOTA methods exhibit significant performance fluctuations depending on the specific dataset and detection granularity (object and point level), indicating they are inherently specialized for narrow sub-domains. In contrast, the proposed framework achieves comprehensive SOTA performance across all metrics and datasets.

4.3 Cross-Task Results

Furthermore, the superiority of DualBP-AD extends beyond achieving the highest object-level detection accuracy and point-level localization precision. While all other baseline approaches are strictly confined to the standard detection paradigm, our framework overcomes this limitation.

To evaluate the versatility of DualBP-AD, we extend our evaluation to the multi-class anomaly detection setting, where a single unified model is trained across all object categories within a dataset. Methods like SeDiR (Kim et al., 2026) require explicit class labels during training for feature clustering. This requirement deviates from the unsupervised paradigm which assumes no label availability. In contrast, our framework operates completely independent of class annotations. Even in this strictly unsupervised setting, DualBP-AD achieves new SOTA results, as detailed in Table 1 and Table 2, respectively. Furthermore, to ensure a strict and fair alignment with the evaluation protocol of SeDiR, we introduce a comparative experiment under the same "learning with class-label" setting as SeDiR, further validating its robust capability across varied task constraints.

We further evaluate the framework's robustness in few-shot anomaly task. While the standard protocols for point cloud datasets inherently represent 4-shot scenarios, we introduce stricter 1-, 2-, and 3-shot configurations. Notably, DualBP-AD surpasses existing single-class baselines using only 2 shots on Anomaly-ShapeNet (92.5%) and 3 shots on Real3D-AD (86.2%). Despite these data constraints, the framework maintains highly competitive detection and localization accuracies.

Table 3: Zero-shot 3D anomaly detection under the within-dataset setting. Each cell reports O-AUROC (P-AUROC) in %.

Method	ShapeNet	Real3D	MVTec3D	Eyecandies
3DzAL	58.6(66.1)	52.4(67.5)	53.1(70.8)	55.8(72.7)
BTP	65.2(87.3)	61.4(84.5)	67.8(91.6)	–
MVP	78.7(70.4)	74.9(75.9)	85.7(96.5)	78.2(94.9)
PointAD	82.6(74.1)	74.8(73.5)	86.9(97.2)	77.8(95.3)
AdaCLIP	83.1(60.7)	74.8(51.2)	67.2(94.5)	71.7(86.8)
GS-CLIP	84.1(75.2)	76.4(76.3)	88.2(97.6)	79.3(95.8)
Ours	88.1(96.4)	71.2(96.0)	80.4(92.3)	76.9(91.2)

Table 4: Ablation studies on back-projection aggregation/dual-path fusion/Gaussian aggregation.

Setting	ShapeNet	Real3DAD
mean S^{2D}	95.0	86.1
geom_S^{2D}	95.0	86.3
ours_S^{2D}	95.3	87.3
mean S^{3D}	94.3	84.1
max S^{3D}	94.7	85.7
ours_S^{3D}	95.0	86.6
S^{fused}	95.3	87.6
G-S^{fused}	**95.6**	**88.3**

Finally, we evaluate zero-shot performance by randomly partitioning dataset categories into distinct training and testing sets. Unlike traditional supervised methods, our framework strictly maintains an unsupervised multi-class paradigm, achieving strong zero-shot capabilities completely without labels. Notably, DualBP-AD attains stable and SOTA results on Anomaly-ShapeNet (see Table 3), demonstrating that our method surpasses traditional approaches given ample category diversity and sample volume. However, on datasets with limited training categories, the detection accuracy becomes highly sensitive to the partition configuration. This reveals that while our label-free model excels with abundant data, its zero-shot generalization under data-scarce conditions currently lags behind traditional vision-language model-based approaches.

4.4 Ablation Study

We evaluate the individual contributions of the dual-path back-projection mechanism, which computes anomaly scores at the 2D pixel level (S^{2D}) and the 3D point level (S^{3D}). Utilizing either path independently yields competitive detection results; however, they exhibit distinct functional biases depending on the anomaly typology and data modality. The S^{2D} demonstrates exceptional sensitivity to high-frequency, localized surface defects (e.g., textural variations or superficial scratches). Conversely, the S^{3D} is superior in identifying low-frequency structural deformations and topological geometric inconsistencies. Consequently, the point-level fusion of these two metrics integrates orthogonal error spaces, enabling the framework to dynamically adapt to diverse defect distributions and achieve optimal generalization across varying dataset complexities.

Furthermore , we ablate the specific aggregation strategies utilized within the back-projection layer. For the score-level aggregation, applying geometric weighting is better than standard mean pooling by effectively filtering perspective artifacts by prioritizing orthogonal camera rays. Furthermore, the top3-geom strategy prevents the dilution of highly localized anomaly signals. For the feature-level aggregation, employing max pooling during inference explicitly preserves localized structural anomalies better than mean pooling. However, mean pooling remains strictly necessary during the training phase to provide dense and stable gradient updates.

5 Conclusion

In this paper, we propose DualBP-AD, the first universal framework for 3D anomaly detection that overcomes the reliance on specific task configurations and single-modality inputs. Extensive experiments demonstrate that our unified model achieves state-of-the-art performance across diverse benchmarks. Practically, this universality eliminates the need for continuous model redeployment, allowing a single framework to adapt to varying and dynamic industrial scenarios.

Admittedly, DualBP-AD is less sensitive to fine-grained anomalies within complex geometric topologies compared to methods operating directly on raw coordinates. Future work will focus on two directions: integrating native geometric priors to enhance topological awareness, and further optimizing the model's adaptability for complex, real-world applications.

AI ASSISTANCE DISCLOSURE

We disclose the use of AI assistants and Large Language Models during the development of this research and the preparation of this paper.

For retrieval and discovery, Claude Sonnet 5, Gemini 3.1, GPT 5.5, and DeepSeek v4 assisted in surveying the research background and identifying relevant related works, with a primary focus on confirming appropriate datasets and baselines for our experiments.

Regarding research execution, we employed Claude Sonnet 5 to assist with software development and data management. It provided technical guidance for debugging our implementation and reducing memory overhead. Furthermore, we used Claude Sonnet 5 locally to aggregate and process experimental results across different datasets and random seeds. This data organization facilitated the systematic analysis of our main experimental results and hyperparameter evaluations.

For writing assistance and manuscript polishing, Claude Sonnet 5 was utilized to generate a foundational structural outline based on our source code and raw experimental outcomes. We emphasize that this AI-generated framework served strictly as a structural reference; no AI-generated prose was directly copied into the paper. Finally, Claude Opus 5 was used to proofread the manuscript for grammatical and LaTeX formatting errors, as well as to provide simulated peer-review feedback.

REFERENCES

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*, pp. 202–213. SCITEPRESS - Science and Technology Publications, 2022. doi: 10.5220/0010865000003124. URL http://dx.doi.org/10.5220/0010865000003124.

Luca Bonfiglioli, Marco Toschi, Davide Silvestri, Nicola Fioraio, and Daniele De Gregorio. The eyecandies dataset for unsupervised multimodal anomaly detection and localization, 2022. URL https://arxiv.org/abs/2210.04570.

J. J. Cabrera, A. Santo, A. Gil, C. Viegas, and L. Payá. Minkunext: Point cloud-based large-scale place recognition using 3d sparse convolutions, 2024. URL https://arxiv.org/abs/2403.07593.

Yunkang Cao, Xiaohao Xu, and Weiming Shen. Complementary pseudo multimodal feature for point cloud anomaly detection. *Pattern Recognition*, 156:110761, 2024a. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2024.110761. URL https://www.sciencedirect.com/science/article/pii/S0031320324005120.

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. *AdaCLIP: Adapting CLIP with Hybrid Learnable Prompts for Zero-Shot Anomaly Detection*, pp. 55–72. Springer Nature Switzerland, 2024b. ISBN 9783031727610. doi: 10.1007/978-3-031-72761-0_4. URL http://dx.doi.org/10.1007/978-3-031-72761-0_4.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. URL https://arxiv.org/abs/2104.14294.

Ruitao Chen, Guoyang Xie, Jiaqi Liu, Jinbao Wang, Ziqi Luo, Jinfan Wang, and Feng Zheng. Easynet: An easy network for 3d industrial anomaly detection, 2023. URL https://arxiv.org/abs/2307.13925.

Jiayi Cheng, Can Gao, Jie Zhou, Jiajun Wen, Tao Dai, and Jinbao Wang. Mc3d-ad: A unified geometry-aware reconstruction model for multi-category 3d anomaly detection. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 837–845. International Joint Conferences on Artificial Intelligence Organization, 8 2025a. doi: 10.24963/ijcai.2025/94. URL https://doi.org/10.24963/ijcai.2025/94. Main Track.

Yuqi Cheng, Yihan Sun, Hui Zhang, Weiming Shen, and Yunkang Cao. Towards high-resolution 3d anomaly detection: A scalable dataset and real-time framework for subtle industrial defects, 2025b. URL https://arxiv.org/abs/2507.07435.

Yuqi Cheng, Yunkang Cao, Guoyang Xie, Zhichao Lu, and Weiming Shen. Toward zero-shot point cloud anomaly detection: A multiview projection framework. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 56(3):1747–1760, 2026. doi: 10.1109/TSMC.2025.3648581.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization, 2020. URL https://arxiv.org/abs/2011.08785.

Zehao Deng, An Liu, and Yan Wang. Gs-clip: Zero-shot 3d anomaly detection by geometry-aware prompt and synergistic view representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35587–35596, June 2026.

Xintong Dong, Chuanyang Li, Peng Zheng, Chuqi Han, Jiaxin Jing, Hailong Shen, Yanzhi Song, and Zhouwang Yang. Histcad: A constraint-aware parametric history-based cad representation, dataset, and benchmark with industrial complexity, 2026. URL https://arxiv.org/abs/2602.19171.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

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, 2019. URL https://arxiv.org/abs/1904.02639.

Lipeng Gu, Xuefeng Yan, Liangliang Nan, Dingkun Zhu, Honghua Chen, Weiming Wang, and Mingqiang Wei. Pointenet: A lightweight framework for effective and efficient point cloud analysis, 2023. URL https://arxiv.org/abs/2312.12743.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Univad: A training-free unified model for few-shot visual anomaly detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15194–15203, 2025. doi: 10.1109/CVPR52734.2025.01415.

Jia Guo, Shuai Lu, Lei Fan, Zelin Li, Donglin Di, Yang Song, Weihang Zhang, Wenbing Zhu, Hong Yan, Fang Chen, Huiqi Li, and Hongen Liao. One dinomaly2 detect them all: A unified framework for full-spectrum unsupervised anomaly detection, 2025a. URL https://arxiv.org/abs/2510.17611.

Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection, 2025b. URL https://arxiv.org/abs/2405.14325.

Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. Diad: A diffusion-based framework for multi-class anomaly detection, 2023. URL https://arxiv.org/abs/2312.06607.

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 (CVPR) Workshops*, pp. 2968–2977, June 2023.

Xueyang Kang, Zizhao Li, Tian Lan, Dong Gong, Kourosh Khoshelham, and Liangliang Nan. Hierarchical point-patch fusion with adaptive patch codebook for 3d shape anomaly detection, 2026. URL https://arxiv.org/abs/2604.03972.

SuYeon Kim, Wongyu Lee, and MyeongAh Cho. A semantically disentangled unified model for multi-category 3d anomaly detection, 2026. URL https://arxiv.org/abs/2603.25159.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. URL https://arxiv.org/abs/2304.02643.

Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. Abc: A big cad model dataset for geometric deep learning, 2019. URL https://arxiv.org/abs/1812.06216.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9664–9674, June 2021.

Kaiqiang Li, Gang Li, Mingle Zhou, Min Li, Delong Han, and Jin Wan. Back to point: Exploring point-language models for zero-shot 3d anomaly detection, 2026. URL https://arxiv.org/abs/2603.21511.

Wenqiao Li, Xiaohao Xu, Yao Gu, Bozhong Zheng, Shenghua Gao, and Yingna Wu. Towards scalable 3d anomaly detection and localization: A benchmark via 3d anomaly synthesis and a self-supervised learning network, 2023. URL https://arxiv.org/abs/2311.14897.

Hanzhe Liang, Aoran Wang, Jie Zhou, Xin Jin, Can Gao, and Jinbao Wang. Examining the source of defects from a mechanical perspective for 3d anomaly detection, 2025. URL https://arxiv.org/abs/2505.05901.

Jiaqi Liu, Guoyang Xie, Ruitao Chen, Xinpeng Li, Jinbao Wang, Yong Liu, Chengjie Wang, and Feng Zheng. Real3d-ad: A dataset of point cloud anomaly detection, 2023a. URL https://arxiv.org/abs/2309.13226.

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20402–20411, June 2023b.

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, 2025a. URL https://arxiv.org/abs/2503.02424.

Wei Luo, Haiming Yao, Yunkang Cao, Qiyu Chen, Ang Gao, Weiming Shen, and Wenyong Yu. Inpformer++: Advancing universal anomaly detection via intrinsic normal prototypes and residual learning, 2025b. URL https://arxiv.org/abs/2506.03660.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning, 2022. URL https://arxiv.org/abs/2203.06604.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. URL https://arxiv.org/abs/2408.00714.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection, 2022. URL https://arxiv.org/abs/2106.08265.

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

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Chengyu Tao, Xuanming Cao, and Juan Du. G^2sf-miad: Geometry-guided score fusion for multimodal industrial anomaly detection, 2025. URL https://arxiv.org/abs/2503.10091.

Chengjie Wang, Haokun Zhu, Jinlong Peng, Yue Wang, Ran Yi, Yunsheng Wu, Lizhuang Ma, and Jiangning Zhang. M3dm-nr: Rgb-3d noisy-resistant industrial anomaly detection via multimodal denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):9981–9993, 2025a. doi: 10.1109/TPAMI.2025.3592089.

Yizhou Wang, Kuan–Chuan Peng, and Yun Fu. Towards zero-shot 3d anomaly localization. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1447–1456, 2025b. doi: 10.1109/WACV61041.2025.00148.

Rundi Wu, Chang Xiao, and Changxi Zheng. Deepcad: A deep generative network for computer-aided design models, 2021. URL https://arxiv.org/abs/2105.09492.

Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling, 2022. URL https://arxiv.org/abs/2210.05666.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger, 2024. URL https://arxiv.org/abs/2312.10035.

Chaoran Xu, Chengkan Lv, Qiyu Chen, Feng Zhang, and Zhengtao Zhang. Mrad: Zero-shot anomaly detection with memory-driven retrieval. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 45751–45784, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4c12e97f2e05304a451e18c9c945036f-Paper-Conference.pdf.

Shaoran Yang, Dongyu Wei, Hanzhi Yu, Zhaohui Yang, Yuchen Liu, and Mingzhe Chen. Contrastive language-image pre-training model based semantic communication performance optimization, 2025. URL https://arxiv.org/abs/2507.08873.

Jianan Ye, Weiguang Zhao, Xi Yang, Guangliang Cheng, and Kaizhu Huang. Po3ad: Predicting point offsets toward better 3d point cloud anomaly detection, 2024. URL https://arxiv.org/abs/2412.12617.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Keep drÆming: Discriminative 3d anomaly detection through anomaly simulation. *Pattern Recognition Letters*, 181:113–119, 2024. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2024.03.018. URL https://www.sciencedirect.com/science/article/pii/S0167865524000862.

Yaohua Zha, Xue Yuerong, Chunlin Fan, Yuansong Wang, Tao Dai, Ke Chen, and Shu-Tao Xia. Casl: Curvature-augmented self-supervised learning for 3d anomaly detection, 2025. URL https://arxiv.org/abs/2511.12909.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer, 2021. URL https://arxiv.org/abs/2012.09164.

Bozhong Zheng, Jinye Gan, Xiaohao Xu, Xintao Chen, Wenqiao Li, Xiaonan Huang, Na Ni, and Yingna Wu. Bridging 3d anomaly localization and repair via high-quality continuous geometric representation, 2025. URL https://arxiv.org/abs/2505.24431.

Qihang Zhou, Jiangtao Yan, Shibo He, Wenchao Meng, and Jiming Chen. Pointad: Comprehending 3d anomalies from points and pixels for zero-shot 3d anomaly detection, 2024. URL https://arxiv.org/abs/2410.00320.

Qihang Zhou, Shibo He, Jiangtao Yan, Wenchao Meng, and Jiming Chen. Pointad+: Learning hierarchical representations for zero-shot 3d anomaly detection, 2025. URL https://arxiv.org/abs/2509.03277.

APPENDIX

A SYMBOLS AND CONFIGURATION

In this section, we detail the symbol definitions and the hyperparameter configurations utilized in our DualBP-AD framework. The complete list of mathematical notations and hyperparameters are summarized in Table 5 and Table 6, respectively. During the forward pass, given an input 3D point cloud $\mathcal{P}$ with N points, the framework first renders $V = 12$ RGB images I_v at a native resolution of 518×518, which are subsequently center-cropped to 512×512 before being fed into the visual encoder. Concurrently, this rendering process establishes the pixel-to-point correspondence $\mathcal{C}_v$, the point visibility mask $\delta_{v,i}$, and the incidence angle $\theta_{v,i}$ for each point p_i, providing essential geometric priors for the subsequent back-projection.

The cropped images are processed by a pre-trained DINOv3 ViT-B encoder to extract robust 2D semantic features $\mathbf{E}_v^l$ across $L = 4$ fused feature levels with a channel dimension of $C = 768$. To resolve contextual ambiguity across varying viewpoints and categories, the instance-adaptive spatial-semantic transition layer employs $M = 6$ intrinsic normal prototypes (INPs), denoted as $\mathbf{k}_m$. This mechanism adaptively aligns the extracted 2D spatial features with explicit 3D semantic representations, thereby producing the refined decoder features $\mathbf{D}_v^l$.

Following the decoding stage, the framework generates a 2D anomaly map $\mathbf{A}_v$ for each view. To accurately map these multi-view 2D observations back into the native 3D topology, we introduce an orthogonal dual-path back-projection mechanism. Specifically, the 2D anomaly scores $a_v(p_i)$ are projected to formulate the score-level 3D anomaly score $s^{\mathrm{2D}}(p_i)$, while the deep feature representations are simultaneously mapped to compute the feature-level score $s^{\mathrm{3D}}(p_i)$. This dual projection is continuously modulated by a geometric reliability weight $W^{\mathrm{geom}}(p_i, v)$—derived from the incidence angle $\theta_{v,i}$—to filter out severely distorted viewing angles. Finally, the dual-path scores are smoothed using a K-nearest neighbor ($\mathcal{N}_K(p_i)$) Gaussian filter with a neighborhood size of $K = 25$ and a bandwidth of $\sigma = 1.0$, culminating in a fused point-level anomaly score and a top-$z\%$ aggregated object-level score.

To ensure stable optimization, the training process is decoupled into two distinct stages. Stage-1 focuses on 2D representation learning over $E_1 = 100$ epochs, optimized primarily by a pixel-level reconstruction loss and a soft INP coherence loss $\mathcal{L}_{\mathrm{sc}}$ weighted by $\lambda_1 = 0.2$. Stage-2 integrates the 3D topology over $E_2 = 50$ epochs, jointly optimizing the unified objective with a 3D loss weight of $\lambda_2 = 0.2$ calculated on $N_s = 4096$ FPS-sampled visible points ($\tilde{\mathcal{P}}_{\mathrm{vis}}$). Furthermore, an error-focusing mining exponent of $\gamma = 3$ is applied throughout the objective function to non-linearly amplify the gradient signals originating from hard anomalous regions.

B SUPPLEMENTARY RESULTS

B.1 ANOMALY-SHAPENET

Anomaly-ShapeNet is a benchmark for 3D anomaly detection comprising 40 distinct object categories. Compared to several conventional industrial datasets, it contains a broader range of geometric variations and topological structures. The intra-class variance and diverse 3D surfaces provide a suitable environment for evaluating the generalizability of anomaly detection models across different shapes.

Due to space constraints in the main text, we report the complete per-category localization results (P-AUROC) on Anomaly-ShapeNet in Table 7. In the single-class setting, existing methods show varying degrees of performance across the 40 categories. Under the multi-class setting, we observe a decrease in the performance of baseline methods such as MC3D-AD and SeDiR, with their mean P-AUROC recorded at 75.9% and 81.1%, respectively. This performance gap may be related to the challenges posed by cross-category topological variations in a unified training setup.

For DualBP-AD, the mean P-AUROC is 97.3% in both single-class and multi-class configurations using the same architecture. This suggests that the proposed instance-adaptive spatial-semantic transition layer can help mitigate multi-category interference to some extent. In the 2-shot setting, the

model achieves a mean P-AUROC of 96.9%, which indicates its potential applicability in scenarios with limited training data.

To evaluate the stability of the proposed method, the reported results are computed as the average of 10 independent experiments using different random seeds. Taking single-class task for example, for the object-level anomaly detection metric (O-AUROC), the model achieves a maximum of 95.9% and a minimum of 95.0%, with a mean variance of 5.6×10^{-6}. For the point-level localization metric (P-AUROC), the performance ranges from a maximum of 97.4% to a minimum of 97.3%, yielding a variance of 5.8×10^{-8}. These low variance values indicate that DualBP-AD remains stable across different random initializations.

Table 5: Symbols used in the main text and formulas.

Symbol	Domain	Meaning
$\mathcal{P}$	point set	Input 3D point cloud
N	integer	Number of points in $\mathcal{P}$
p_i	$\mathbb{R}^3$	i-th point
$\mathbf{n}_i$	$\mathbb{R}^3$	Surface normal at p_i
V	integer	Number of rendered views (12)
I_v	image	Rendered RGB image of view v
$\mathbf{E}_v^{\text{cam}}$	4×4	Extrinsic matrix of view v
$\mathcal{C}_v$	map	Pixel-to-point correspondence of view v
$u_{v,i}, w_{v,i}$	pixels	Projected coordinates of p_i in view v
$\delta_{v,i}$	$\{0, 1\}$	Visibility of p_i in view v
$\mathcal{V}(i)$	set	Visible views of p_i
$\theta_{v,i}$	$[0, \pi]$	Incidence angle, normal vs. viewing ray
$\mathbf{E}_v^l$ / $\mathbf{D}_v^l$	feature maps	Encoder / decoder features, level l
L	integer	Number of fused feature levels
$\mathbf{f}_n$ / $\mathbf{f}_{v,n}^l$	vector	Patch token (view v, position n, level l)
$\mathbf{d}_{v,n}^l$	vector	Reconstructed token
$\mathbf{Q}$ / $\mathbf{K}, \mathbf{V}$	matrices	Decoder query / key-value inputs
$\mathbf{k}_m$	vector	m-th intrinsic normal prototype (INP)
M	integer	Number of INPs (6)
C	integer	Feature channels (768)
$\mathbf{A}_v$	map	Per-view 2D anomaly map
$a_v(p_i)$	scalar	2D anomaly score of p_i in view v
$W^{\text{geom}}(p_i, v)$	scalar	Geometric reliability weight
δ_{floor} / γ	scalars	Weight floor / sharpness
$\mathcal{T}_s(i)$	set	Top-s scoring visible views of p_i
$s^{\text{2D}}(p_i)$ / $s^{\text{3D}}(p_i)$	scalars	Score-level / feature-level 3D scores
$\text{pool}(\cdot)$	operator	Cross-view pooling (max at inference, mean in training)
$w_{i,j}^{\text{gauss}}$	scalar	Gaussian neighbor weight
σ / K	scalars	Smoothing bandwidth / neighborhood size (1.0 / 25)
$\mathcal{N}_K(p_i)$	set	K-nearest neighbors of p_i
$\mathcal{X}$ / $\hat{\mathcal{X}}$	generic	Property before / after smoothing
$\hat{S}^{\text{2D}}, \hat{S}^{\text{3D}}, \hat{S}^{\text{fused}}$	scalars	Branch and fused object scores
z	percent	Top-$z\%$ aggregation ratio
α	$[0, 1]$	2D-3D fusion weight
$\mathcal{L}_{\text{sc}}$	scalar	Soft INP coherence loss
$\mathcal{L}_{\text{recon}}^{\text{2D}}$ / $\mathcal{L}_{\text{recon}}^{\text{3D}}$	scalars	Pixel-level / point-level reconstruction losses
μ_{2D} / γ_{2D}	scalars	2D error mean / focusing exponent
$\nu(p_i)$ / η	scalars	Local normal variance / scaling factor
$\tilde{\mathcal{P}}_{\text{vis}}$	set	Visible FPS-sampled points
$\mathcal{L}_{\text{2D}}$ / $\mathcal{L}_{\text{3D}}$	scalars	Mined 2D / 3D training losses
$\mathcal{L}_{\text{pre}}$ / $\mathcal{L}_{\text{post}}$	scalars	Stage-1 / stage-2 objectives
E_1 / E_2	integers	Stage epochs (100 / 50)
λ_1 / λ_2	scalars	Coherence / 3D loss weights (0.2 / 0.2)

Table 6: Hyperparameter summary.

Group	Param	Value
Rendering	V_i (views)	12 (default)
	Native render $H_0 \times W_0$	518×518
	Model input $H \times W$ (center crop)	512×512
	Feature map $h \times w$ ($H/16$)	32×32
Encoder	Backbone	DINOv3 ViT-B
	Channels C	768
	Fused levels L	2
	Features	$\mathbf{E}^{\ell}_{i,v}, \mathbf{D}^{\ell}_{i,v} \in \mathbb{R}^{C \times h \times w}$
INP	Number of INPs M	6
	Coherence weight β	0.2
Training	Stage-1 epochs E_1 / Stage-2 epochs E_2	100 / 50
	3D loss weight λ_{3D}	0.2
	Mining exponent γ	3
	FPS points for $\mathcal{L}_{\text{3D}}$ (N_s)	4096
	Stage-1 $\mathcal{L}_{\text{2D}} = \mathcal{L}_{\text{rec}} + \beta \mathcal{L}_{\text{coh}}$	0.1
	Stage-2 $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{2D}} + \lambda_{\text{3D}} \mathcal{L}_{\text{3D}}$	0.1
Inference	Score fusion $s^{\text{fuse}}_{i,p} = \alpha \mathcal{N}_f(s^f) + (1-\alpha)\mathcal{N}_s(s^s)$	$\alpha \in [0,1]$
	Object score	top-z% aggregation $S_i(z)$

Method	ashtray0	bag0	bottle0	bottle1	bottle3	bowl0	bowl1	bowl2	bowl3	bowl4	bowl5	bucket0	bucket1	cap0	cap3	cap4	cap5	cup0	cup1	eraser0	Mean
Single-class Method																					
PO3AD(CVPR'25)	96.2	94.9	91.2	84.4	88.0	97.8	91.4	91.8	93.5	96.7	94.1	75.5	89.9	95.7	94.8	94.0	86.4	90.9	93.2	97.4	**89.9**
Simple3D(AAAI'26)	92.0	95.4	97.4	72.8	83.8	98.8	95.1	93.3	99.3	92.9	97.9	72.5	92.1	98.8	96.4	97.9	96.4	97.9	93.7	97.0	**92.9**
CASL(AAAI'26)	88.7	96.8	85.3	86.2	91.4	95.8	97.5	98.2	99.8	99.6	97.4	79.6	92.7	97.4	97.1	94.9	66.9	98.0	79.7	87.6	**89.9**
PASDF(ICCV'25)	91.9	95.8	95.1	92.6	94.8	96.3	90.0	81.6	93.9	86.5	90.9	87.5	82.4	94.8	86.1	89.4	92.0	94.8	88.4	94.5	**89.6**
MC4AD(arXiv'25)	90.4	95.7	94.2	88.1	89.6	97.7	92.1	96.5	93.7	99.3	95.2	79.2	92.5	95.9	94.6	94.0	92.5	91.7	92.0	95.1	**91.0**
Ours(2-shot)	97.1	96.5	99.4	98.9	94.2	93.7	98.1	99.4	96.8	98.5	97.8	97.6	96.8	97.6	98.8	98.9	98.8	95.2	95.8	96.6	**96.9**
Ours	97.5	96.4	99.7	99.3	95.1	94.4	98.2	99.4	96.9	98.7	98.0	97.6	97.0	98.2	99.2	99.0	99.1	96.8	96.5	96.6	**97.3**
Multi-class Method																					
MC3D-AD(IJCAI'25)	80.1	81.5	89.5	88.3	90.1	82.7	53.4	60.7	78.3	65.4	55.2	79.0	89.5	86.7	92.2	88.0	88.5	82.9	71.5	85.6	**75.9**
SeDiR(CVPR'26)	76.4	89.8	91.1	89.8	94.5	86.6	70.0	79.5	86.9	80.6	64.6	71.7	88.4	90.0	97.2	94.0	95.0	87.1	84.2	74.8	**81.1**
Ours	96.7	96.4	99.7	99.3	95.1	94.6	98.5	99.6	97.2	98.9	98.4	97.2	96.7	97.8	99.4	99.0	99.3	96.8	96.2	96.7	**97.3**
Ours(label)	97.4	96.6	99.7	99.4	95.2	94.9	98.6	99.6	97.4	99.0	98.5	97.5	96.9	97.9	99.5	99.1	99.4	96.9	96.3	96.7	**97.5**

Method	headset0	headset1	helmet0	helmet1	helmet2	helmet3	jar	phone	shelf0	tap0	tap1	vase0	vase1	vase2	vase3	vase4	vase5	vase7	vase8	vase9	Mean
Single-class Method																					
PO3AD(CVPR'25)	82.3	90.7	87.8	94.8	93.2	84.6	87.1	81.0	66.3	78.3	69.2	95.5	88.2	97.8	88.4	90.2	93.7	98.2	95.0	95.2	**89.9**
Simple3D(AAAI'26)	93.6	95.8	90.2	90.2	94.7	92.8	97.8	96.4	90.1	85.7	81.1	93.4	80.7	98.6	91.0	98.5	96.6	99.0	97.5	91.3	**92.9**
CASL(AAAI'26)	77.0	90.9	92.3	88.2	92.9	98.0	97.4	83.8	92.8	65.2	60.2	89.4	97.2	97.1	93.1	91.3	73.2	99.7	98.9	88.6	**89.9**
PASDF(ICCV'25)	86.3	89.1	81.6	64.6	80.9	95.8	95.9	95.1	86.5	88.4	90.2	94.4	79.7	95.6	86.8	89.9	91.5	95.9	90.9	86.3	**89.6**
MC4AD(arXiv'25)	78.0	93.2	88.5	92.6	87.4	87.9	87.7	83.3	81.0	79.8	77.7	96.9	88.9	99.0	88.2	87.5	93.3	99.5	95.3	95.9	**91.0**
Ours(2-shot)	98.6	96.3	97.3	93.7	95.9	92.8	93.5	99.2	98.2	97.8	97.0	99.3	95.0	92.4	96.0	99.1	96.1	96.4	97.0	98.4	**96.9**
Ours	98.9	96.6	97.5	94.0	96.1	93.1	94.4	99.3	98.3	98.9	97.4	99.5	95.9	92.7	96.8	99.3	96.7	97.2	97.3	98.9	**97.3**
Multi-class Method																					
MC3D-AD(IJCAI'25)	66.6	70.9	74.4	58.2	81.7	60.2	84.8	85.4	65.8	52.4	53.7	87.0	72.6	82.2	79.7	78.6	64.7	61.6	88.5	77.6	**75.9**
SeDiR(CVPR'26)	70.8	68.7	81.0	59.6	88.5	70.1	91.1	89.1	67.9	59.7	63.0	84.6	72.2	85.0	83.0	88.4	76.6	77.7	90.0	83.5	**81.1**
Ours	98.9	96.2	96.1	93.6	96.2	93.2	94.3	99.3	97.6	98.3	97.7	99.5	96.1	93.7	97.2	99.4	96.9	97.4	97.5	99.0	**97.3**
Ours(label)	99.0	96.5	96.5	94.0	96.2	93.5	94.5	99.4	97.9	98.5	97.8	99.6	96.3	93.8	97.4	99.4	97.2	97.6	97.5	99.1	**97.5**

Table 7: Anomaly detection results on Anomaly-ShapeNet (40 categories). Each cell reports P-AUROC in %. Methods are grouped into single-class and multi-class settings.

B.2 REAL3D-AD

Real3D-AD is a dataset designed for high-resolution 3D anomaly detection in real-world industrial scenarios. It comprises point clouds captured by 3D scanners, featuring practical challenges such as sensor noise, incomplete surfaces, and fine-grained structural defects. This dataset serves to evaluate the robustness and applicability of anomaly detection models in authentic manufacturing environments.

Similar to the evaluation on Anomaly-ShapeNet, we assess the stability of DualBP-AD on the Real3D-AD dataset by reporting the average results of 10 independent experiments with different random seeds. For the object-level anomaly detection metric (O-AUROC), the model records a maximum of 89.0% and a minimum of 87.8%, with a mean variance of 1.47×10^{-5}. Regarding the point-level localization metric (P-AUROC), the performance shows a maximum of 96.6% and a minimum of 96.5%, resulting in a variance of 6.3×10^{-8}. These measurements indicate that the

proposed framework maintains consistent performance when processing complex, real-world point cloud data.

Unlike synthetic datasets, Real3D-AD contains real-world partial scans. This causes practical issues like incomplete surfaces and noisy normals. When rendering these partial 3D scans to 2D images, some bad viewpoints (such as back views) can capture dark or missing areas, leading to false anomaly scores. To keep DualBP-AD universal, we do not change the training process. All training process remain identical to those used for Anomaly-ShapeNet. Instead, we only apply a simple view-aggregation strategy during inference. We use basic geometric and appearance rules to filter out unreliable views and only keep the best ones for feature fusion. Finally, we apply spatial smoothing to the anomaly scores to reduce noise near the broken boundaries.

B.3 MVTEC 3D-AD

MVTec 3D-AD is a standard benchmark containing 10 object categories, widely used to evaluate 3D anomaly detection models. Notably, this dataset natively provides highly aligned multi-modal inputs, including RGB images, depth maps, and xyz maps. Given that the paired depth and RGB maps inherently establish a precise dense 2D-3D correspondence, explicit multi-view rendering and back-projection operations provide limited additional performance gains on this specific dataset. To maximize computational efficiency, we introduce a streamlined configuration that bypasses these modules and directly utilizes the native modality alignment. Furthermore, because the 3D geometric constraints are already well-represented in the provided depth maps, the first-stage learning is sufficient for the model to converge.

Table 8 shows the detailed per-class results on the MVTec 3D-AD dataset. We report both object-level (O-AUROC) and point-level (P-AUROC) metrics under single-class and multi-class settings. In the single-class setup, DualBP-AD achieves a mean O-AUROC of 97.4% and a mean P-AUROC of 99.0%. Under the multi-class setup, the model achieves a mean O-AUROC of 95.7% and a mean P-AUROC of 98.9%. These results confirm that the proposed framework performs consistently on standard industrial test sets.

Table 8: Single-class/Multi-class O-AUROC and P-AUROC on MVTec 3D-AD .

Method	Bagel	Cable gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
single-class(O-AUROC)	99.1	98.6	98.0	100.0	98.7	90.9	98.9	97.1	99.0	93.3	**97.4**
single-class(P-AUROC)	99.0	99.5	99.8	99.2	99.7	95.9	99.9	99.9	97.5	99.2	**99.0**
multi-class(O-AUROC)	97.4	97.5	96.0	100.0	95.7	88.1	95.9	94.3	99.8	92.5	**95.7**
multi-class(P-AUROC)	98.8	99.6	99.9	97.6	99.8	95.9	99.7	99.8	99.0	99.2	**98.9**

B.4 EYECANDIES

The Eyecandies dataset provides high-quality multi-modal data focusing on natural and synthetic confectionery products. Specifically, for each sample, it provides six RGB images captured under varying illumination conditions, alongside a corresponding depth map and normal map. All these multi-modal inputs share identical camera intrinsic and extrinsic parameters, ensuring perfect spatial alignment. Similar to the setup used for MVTec 3D-AD, we leverage this native alignment by directly utilizing the provided depth maps. This bypasses the need for explicit multi-view rendering and optimizes computational efficiency.

The detailed per-class performance on the Eyecandies dataset is presented in Table 9. Under the single-class configuration, DualBP-AD records a mean O-AUROC of 95.3% and a mean P-AUROC of 98.9%. When tested in the more challenging multi-class scenario with the identical model architecture and settings, the model maintains a strong performance, achieving a mean O-AUROC of 94.1% and a mean P-AUROC of 98.9%. These consistent results highlight the robustness of the framework across diverse object types.

B.5 ZERO-SHOT

Traditional zero-shot anomaly detection in the 3D domain often relies on a supervised paradigm, where models learn the semantic differences between normal and anomalous samples on a limited

Table 9: Single-class/Multi-class O-AUROC and P-AUROC on Eyecandies.

Method	CandyCane	Choc.Cookie	Choc.Praline	Confetto	GummyBear	Hazelnut	Licorice	Lollipop	Marshmallow	Peppermint	Mean
single-class(O-AUROC)	87.0	99.5	99.0	100.0	93.6	83.7	99.0	91.2	100.0	100.0	**95.3**
single-class(P-AUROC)	99.0	99.8	99.7	99.8	96.8	96.5	99.4	98.3	99.9	99.5	**98.9**
multi-class(O-AUROC)	81.9	99.2	98.6	100.0	93.6	80.5	99.0	87.7	100.0	100.0	**94.1**
multi-class(P-AUROC)	98.6	99.8	99.7	99.9	97.0	96.5	99.6	98.7	99.9	99.7	**98.9**

set of categories before being evaluated on unseen categories. In contrast, DualBP-AD strictly adheres to an unsupervised learning paradigm. To maintain this standard while establishing a valid zero-shot evaluation, we configure the task as follows: we randomly partition the dataset, selecting half of the categories for training and reserving the remaining half for testing. Additionally, a portion of the normal test samples from the training categories is incorporated into the training set. Since there is strictly no intersection between the training and testing categories, and no anomalous data or labels are introduced during training, this setup rigorously respects the unsupervised zero-shot condition.

Experimental observations reveal that the zero-shot generalization of our framework is closely correlated with the scale and diversity of the training data. On Anomaly-ShapeNet, which provides abundant sample volumes and a wide variety of categories, the model demonstrates high stability and achieves state-of-the-art zero-shot performance. On datasets with adequate sample volumes but limited category diversity, such as MVTec 3D-AD and Eyecandies, the model maintains stable scoring patterns, though the overall performance is not optimal. Conversely, on Real3D-AD, where both sample data and category diversity are scarce, the zero-shot performance is noticeably restricted.

Achieving zero-shot anomaly detection under strictly unsupervised conditions holds significant practical value, as it eliminates the reliance on costly anomaly annotations and facilitates rapid model deployment in dynamic industrial environments. However, our experimental results also reveal certain limitations. Compared to traditional zero-shot approaches built upon Vision-Language Models (VLMs) or Large Language Models (LLMs), our label-free framework currently requires a broader diversity of training categories to develop robust generalization capabilities for unseen topologies. Enhancing the zero-shot adaptability of this unsupervised paradigm under data-scarce conditions remains a primary focus for future optimization.

B.6 Few-shot

To further evaluate the data efficiency of DualBP-AD, we conduct few-shot experiments on the Anomaly-ShapeNet and Real3D-AD datasets. Figure 5 illustrates the object-level (O-AUROC) and point-level (P-AUROC) performance under 1-shot, 2-shot, and 3-shot settings, compared with the full training set (Full). For point-level anomaly localization (P-AUROC), the model exhibits remarkable stability, maintaining a consistently high performance exceeding 96% across all settings on both datasets. This indicates that our dual-path back-projection mechanism and 2D foundational features can effectively localize anomalous regions even with very limited spatial priors. For object-level anomaly detection (O-AUROC), the performance demonstrates a steady upward trend as the number of training samples increases. These results validate that DualBP-AD is adaptable to data-scarce industrial scenarios while retaining the capacity to scale positively with additional data.

C Ablation Studies

C.1 Supplementary Ablation Studies

To justify our decoupled optimization strategy, we ablate the training schedule in Table 10. Establishing strong 2D semantic representations via 100 epochs of pre-training yields a solid baseline. Subsequently, applying 50 epochs of 3D topological fine-tuning optimally aligns these extracted features with the native spatial geometry, producing the best performance (e.g., 95.6% O-AUROC on Anomaly-ShapeNet). Conversely, merely extending the 2D pre-training to 150 epochs provides negligible gains. Furthermore, excessively prolonging the 3D fine-tuning phase to 100 epochs leads to a slight performance degradation, indicating potential overfitting to the constrained 3D geometric structures at the expense of generalized 2D semantic power. Therefore, the 100-50 epoch schedule is adopted to maintain a perfect balance.

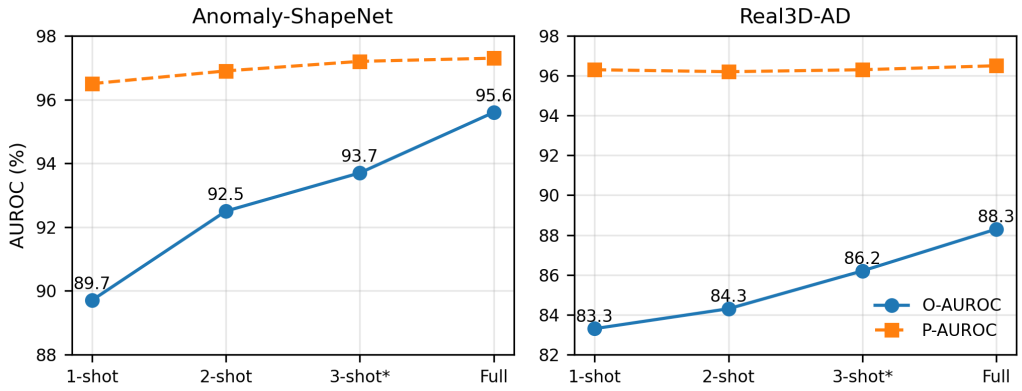


Figure 5: Few-shot performance (O-AUROC and P-AUROC in %) on Anomaly-ShapeNet (left) and Real3D-AD (right).

Table 10: Ablation on the two-stage training schedule (O-AUROC and P-AUROC in %).

Setting	ShapeNet O	ShapeNet P	Real3D O	Real3D P
Pre-train 100	95.2	97.0	87.6	96.2
Pre-train 150	95.3	97.0	87.6	96.1
+ Fine-tune 50 (adopted)	**95.6**	**97.3**	**88.3**	**96.5**
+ Fine-tune 100	95.3	97.1	87.9	96.3

Table 11 demonstrates the efficacy of the instance-adaptive intrinsic normal prototypes. Compared to using static queries, which fail to accommodate massive cross-category topological variations, employing dynamic queries consistently boosts the detection accuracy across diverse datasets. This dynamic configuration enables the spatial-semantic transition layer to adaptively resolve contextual confusion and extract instance-specific geometric features, validating its necessity in our universal multi-class framework.

Table 11: Ablation on the prototype query design (O-AUROC in %).

Query	ShapeNet O	Real3D O
Static queries	94.8	86.9
Dynamic queries	**95.6**	**88.3**

Finally, we evaluate the individual contributions of the error-focusing soft-mining components (Table 12). Without any mining mechanism, the framework treats all regions equally, which may dilute the gradient signals from subtle or localized defects. Introducing 2D mining strictly amplifies the focus on high-frequency superficial anomalies at the pixel level. Further incorporating 3D mining enhances the model's sensitivity to low-frequency structural deformations in the point cloud space.

This section proves the first-order optimization equivalence between our explicit stop-gradient loss reweighting and the bottom-up gradient intervention used in prior works (e.g., INP-Former).

Let $d_i(\theta)$ denote the reconstruction error for the i-th element parameterized by network weights θ, and let $w_i = (d_i/\mu)^\gamma$ be the corresponding difficulty-aware dynamic weight.

For the gradient intervention approach, the forward pass computes the unweighted mean loss $\mathcal{L}_{\text{orig}} = \frac{1}{N}\sum_i d_i(\theta)$. A backward hook subsequently scales the gradient by w_i, producing the final parameter gradient:

$$\frac{\partial \mathcal{L}_{\text{orig}}}{\partial \theta} = \frac{1}{N}\sum_i w_i \frac{\partial d_i(\theta)}{\partial \theta} \tag{14}$$

Our explicit formulation directly computes the weighted loss $\mathcal{L}_{\text{ours}} = \frac{1}{N}\sum_i w_i \cdot d_i(\theta)$. By enforcing a stop-gradient operation on the weight ($\frac{\partial w_i}{\partial \theta} = 0$), the parameter gradient computation bypasses the standard product rule expansion and resolves directly to:

$$\frac{\partial \mathcal{L}_{\text{ours}}}{\partial \theta} = \frac{1}{N}\sum_i w_i \frac{\partial d_i(\theta)}{\partial \theta} \tag{15}$$

Since $\frac{\partial \mathcal{L}_{\text{orig}}}{\partial \theta} \equiv \frac{\partial \mathcal{L}_{\text{ours}}}{\partial \theta}$, our explicit stop-gradient method achieves identical optimization trajectories while seamlessly integrating into standard computational graph frameworks without ad-hoc interventions.

Table 12: Ablation on the soft-mining components (O-AUROC in %).

Setting	ShapeNet O	Real3D O
No mining	95.1	87.4
2D mining	95.4	87.9
2D + 3D mining	**95.6**	**88.3**

Table 13 ablates the K-nearest neighbor Gaussian smoothing on 3D anomaly scores. Without smoothing, localized noise near incomplete boundaries degrades performance. A moderate setting of $K = 25$ and $\sigma = 1.0$ effectively suppresses this noise while preserving anomaly boundaries. Conversely, excessive smoothing $(50, 2.0)$ dilutes localized anomaly signals, leading to decreased accuracy. Thus, $(25, 1.0)$ is adopted.

Table 13: Gaussian smoothing.

(K, σ)	ShapeNet	Real3D
No smoothing	95.3	87.6
$(25, 1.0)$ (adopted)	95.6	88.3
$(50, 2.0)$	95.4	88.0

Figure 6 ablates the impact of 2D rendering resolution and different foundational encoders (MUM, DINOV2-B, and DINOV3-B) on anomaly detection performance. As the rendering size increases from 224 to 512, the O-AUROC for all models improves significantly. To ensure a fair comparison with DINOV2-B, which natively processes 518×518 images, the 512×512 input for DINOV3-B is obtained by center-cropping the original 518-resolution render. Scaling further to 896 yields marginal gains, which does not justify the increased computational overhead. Moreover, DINOV3-B consistently outperforms DINOV2-B and MUM across all resolutions.

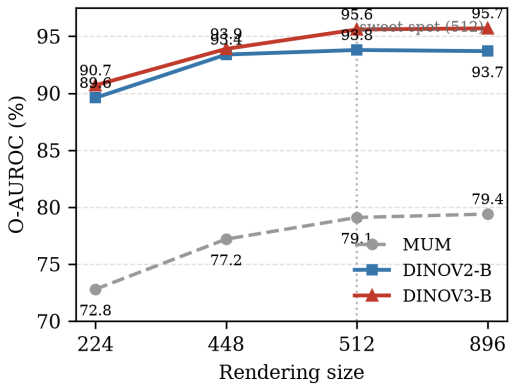


Figure 6: Ablation on rendering size and foundational encoders.