Scientific Reports

https://doi.org/10.1038/s41598-026-41255-2

Article in Press

Efficient industrial point cloud anomaly detection via spatial context aggregation and selective anomalous feature generation

Received: 15 July 2025

Accepted: 18 February 2026

Published online: 24 February 2026

Cite this article as: Hoang D., Tan P.X., Nguyen A. *et al.* Efficient industrial point cloud anomaly detection via spatial context aggregation and selective anomalous feature generation. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-41255-2

Dinh-Cuong Hoang, Phan Xuan Tan, Anh-Nhat Nguyen, Minhhuy Le, Ta Huu Anh Duong, Tuan-Minh Huynh, Duc-Manh Nguyen, Minh-Duc Cao, Duc-Huy Ngo, Minh-Quang Vu, Thu-Uyen Nguyen, Khanh-Toan Phan, Minh-Quang Do, Xuan-Tung Dinh, Van-Hiep Duong & Van-Thiep Nguyen

Efficient Industrial Point Cloud Anomaly Detection via Spatial Context Aggregation and Selective Anomalous Feature Generation

Dinh-Cuong Hoang[1,*], Phan Xuan Tan[2,*], Anh-Nhat Nguyen[3], Minhhuy Le[4], Ta Huu Anh Duong[1], Tuan-Minh Huynh[1], Duc-Manh Nguyen[1], Minh-Duc Cao[1], Duc-Huy Ngo[1], Minh-Quang Vu[1], Thu-Uyen Nguyen[3], Khanh-Toan Phan[3], Minh-Quang Do[3], Xuan-Tung Dinh[3], Van-Hiep Duong[3], and Van-Thiep Nguyen[3]

[1]Greenwich Vietnam, FPT University, Hanoi, 10000, Vietnam
[2]College of Engineering, Shibaura Institute of Technology, Tokyo, 135-8548, Japan
[3]ICT Department, FPT University, Hanoi, 10000, Vietnam
[4]Faculty of Electrical and Electronic Engineering, School of Engineering, Phenikaa University, Hanoi 12116, Vietnam
[*]cuonghd12@fe.edu.vn, tanpx@shibaura-it.ac.jp

ABSTRACT

Automated detection of surface defects on three-dimensional (3D) parts is vital for ensuring product quality and safety in manufacturing. However, three key challenges hinder reliable detection: geometric context ambiguity across complex part shapes, domain mismatch between generic pretrained features and industrial scans (with their unique noise and reflectivity), and the scarcity of diverse defect examples for training. To overcome these issues, we propose a novel single-forward-pass framework for point cloud anomaly detection, comprising three new modules: (1) Spatial Context Aggregation, which grounds each local patch in a set of learned global prototypes via an optimal-transport alignment to resolve context ambiguity; (2) Feature Adaptor, a lightweight two-layer multilayer perceptron (MLP) that fine-tunes self-supervised Point-MAE embeddings to the specific characteristics of industrial scans; and (3) Selective Anomalous Feature Generator, which synthesizes realistic hard negatives by corrupting random subsets of feature tokens, thus mitigating the need for extensive defect labels. An attention-based discriminator trained with patch-wise supervision learns to distinguish these hard negatives from genuine defect-free patterns. At inference, our pipeline delivers dense per-point anomaly scores in a single pass at up to 13.5 frames per second (FPS). On the Real3D-AD benchmark, we observe point-level improvements of 2.8% in area under the receiver operating characteristic curve (AUROC) and 5.7% in area under the precision-recall curve (AUPR), with object-level gains of 3.0% (AUROC) and 3.5% (AUPR). Evaluated on our newly released Industrial3D-AD dataset, which captures realistic sensor noise and reflective materials, we see similar enhancements (2.9%/5.3% point-level, 2.8%/3.3% object-level).

Introduction

In modern manufacturing lines, even minute surface defects can propagate into catastrophic failures. These range from microscopic cracks in aerospace components to subtle deformations in precision-molded parts[1,2]. Such defects can inflict millions of dollars in recall costs and may endanger human lives[3–6]. Traditional two-dimensional inspection systems, though widespread, often miss these fine-scale anomalies when they occur on complex geometries or under variable lighting. As a result, critical vulnerabilities can remain undetected[7–10].

Recent advances in unsupervised image-based anomaly detection have achieved remarkable accuracy by modeling normal appearance patterns and flagging deviations via reconstruction errors or pretrained feature embeddings[11–14]. However, these methods are fundamentally constrained by their two-dimensional nature. Occlusions or specular highlights can obscure defects, and critical geometric cues are lost in projection. As a result, surface irregularities that alter only the third dimension may evade detection altogether. Point cloud anomaly detection addresses these limitations by directly capturing the full three-dimensional shape of inspected objects. Early registration pipelines using handcrafted descriptors established the viability of 3D methods[15]. Subsequent unsupervised frameworks have applied student-teacher distillation to local 3D descriptors[16], constructed memory banks of dual features[17], and exploited group-level feature alignment[18]. Iterative reconstruction frameworks such as IMRNet[19] and diffusion-based restorers like R3D-AD[20] have further enhanced localization accuracy. Despite these advances, most existing approaches incur substantial memory or computational cost, rely on synthetic negatives that may not reflect real defect diversity, and can struggle to generalize across varied geometries and sensor noise patterns.

To address these challenges, we introduce a lightweight and single-pass framework that directly tackles three key obstacles in 3D anomaly detection. The first is the ambiguity in interpreting local geometric deviations without global context. The second is the distributional shift between pretrained features and noisy industrial scans. The third is the scarcity of diverse defect examples needed for training robust detectors. Our approach builds on a Masked Autoencoder pretrained on ShapeNet and incorporates three innovative modules.

First, a Spatial Context Aggregation mechanism uses optimal-transport-based prototype matching to anchor local patch features within a learned global shape context, reducing false positives arising from normal geometric variation. Second, a compact two-layer multilayer perceptron (MLP) serves as a Feature Adaptor that refines generic Point-MAE embeddings to align with the distinct statistics of industrial scans, thereby mitigating domain mismatch with minimal computational overhead. Third, to simulate a broad spectrum of potential anomalies without relying on manually labeled defects, our Selective Anomalous Feature Generator injects scaled Gaussian noise into random subsets of adapted feature tokens, producing realistic hard negatives in feature space. A lightweight attention-based discriminator, trained under patch-wise binary cross-entropy, learns to discriminate these synthetic anomalies from authentic defect-free patterns.

We evaluate our framework on two complementary datasets. Real3D-AD[17] provides ultra-high-precision scans of twelve object classes with carefully annotated bulges and sinks, serving as a controlled benchmark. Our newly released Industrial3D-AD dataset captures ten diverse real-world parts under actual factory conditions, complete with sensor quantization noise, partial occlusions, and subtle surface imperfections. Across both datasets, our method achieves state-of-the-art point- and object-level AUROC and AUPR while maintaining real-time throughput above 13 frames per second. These results demonstrate both improved detection performance and practical deployment readiness.

In summary, our contributions are:

- A unified, single-pass 3D anomaly detection framework that integrates Spatial Context Aggregation, Feature Adaptor, and a selective Anomalous Feature Generator with an attention-based discriminator.
- Spatial Context Aggregation, a parameter-free module that fuses local geometric neighborhoods with global prototypes via optimal transport to sharpen fine-scale defect cues.
- A lightweight Feature Adaptor that fine-tunes pretrained Masked Autoencoder features to industrial domains, aligning feature distributions with minimal computational cost.
- Anomaly synthesis and detection via selective feature-space corruption and a cross-patch attention discriminator trained with patch-wise binary supervision.
- Comprehensive evaluation on Real3D-AD and Industrial3D-AD demonstrating state-of-the-art point- and object-level AUROC/AUPR and over 13 FPS inference speed.

Related Work

Industrial Image Anomaly Detection

Industrial image anomaly detection (IAD) can be organized by the level of supervision during training: supervised[21,22], semi-supervised[23,24], and unsupervised methods[11,12,25,26]. Supervised IAD treats defect identification as a standard classification task, requiring comprehensive labels for both normal and anomalous samples. Semi-supervised approaches leverage a limited set of anomaly labels alongside abundant normal data, balancing annotation effort against performance. Unsupervised methods, trained solely on defect-free images, detect anomalies by modeling normal patterns and flagging deviations at test time.

Supervised IAD approaches employ conventional classifiers such as convolutional neural networks or vision transformers to distinguish between normal and defect classes when enough labeled anomalies are available. For example, attention-based CNNs focus on salient defect regions[21], while transformer architectures incorporate global context for detecting irregularities[22]. Though these methods can achieve high accuracy with diverse anomaly annotations, their deployment is limited by the difficulty of collecting comprehensive labeled datasets spanning all potential defect types and variations. Semi-supervised methods bridge the gap by combining a small number of labeled anomalies with a larger set of defect-free images. Techniques include pseudo-labeling of unlabeled data, modifications to classification loss functions to account for class imbalance, and integration of unsupervised representations with supervised classifiers. Neural network adaptations learn from sparse anomaly labels[23] and balance representation learning on normal data with discriminative fine-tuning on known anomalies[24]. While these techniques can outperform purely unsupervised models when anomaly labels are available, performance still hinges on the quality and diversity of labeled samples and can degrade if anomalies are too heterogeneous.

Unsupervised IAD dispenses with anomalous labels altogether, training exclusively on defect-free images and flagging deviations at inference. Reconstruction-based IAD trains generative models using autoencoders, GANs, diffusion, and transformer

networks solely on normal data under the hypothesis that anomalies yield higher reconstruction errors. Autoencoder variants include multi-scale regional feature guidance[12], divide-and-assemble block-wise memory[27], joint reconstruction-discrimination embeddings[28], and dual subspace re-projection (DSR)[29] that perturbs quantized features to simulate near-in-distribution defects. Data-scarce scenarios are addressed by few-shot reconstruction frameworks[30] and synthetic anomaly generation networks[31]. GAN-based pipelines incorporate adversarial losses with contextual masking or frequency decoupling[32–34]. Emerging diffusion and transformer models include masked transformers for context-aware inpainting[35], dual attention frameworks[36], adaptive inpainting networks (AMI-Net)[37], and dynamic diffusion models that produce high-fidelity reconstructions and pixel-wise anomaly scores[38–40]. Though versatile, these methods demand careful tuning to prevent over-smoothing or reconstructing anomalies. Feature embedding-based approaches leverage pretrained networks to extract semantic representations of normal images and measure discrepancies at test time. Teacher-student distillation frameworks train a student network to imitate a fixed teacher's embeddings, with large divergences signaling defects[11,25,26,41,42]. Memory-bank models store prototypical feature descriptors of normal images and detect anomalies by nearest-neighbor or probabilistic distance metrics[13,43–46]. However, the domain shift between industrial data and pretraining datasets like ImageNet often leads to feature mismatch. Synthesis-based approaches[28,47] aim to generate anomalies on top of clean images.[28] trained discriminatively using synthetically generated near-OOD patterns, while[47] introduces a simple yet effective method that cuts a patch from an image and pastes it elsewhere to simulate anomalies. A CNN is then trained to distinguish between original and augmented distributions. However, such synthetic anomalies rarely match the visual characteristics of real defects, and due to the unpredictable nature of anomalies, exhaustively modeling all possible outliers remains infeasible. To overcome these limitations,[48] introduced a feature adaptor that fine-tunes pretrained CNNs on the target domain, reducing feature bias. Instead of synthesizing anomalies in pixel space, SimpleNet operates in feature space, improving realism and relevance. It uses a single-stream architecture at inference and is entirely composed of conventional CNN blocks, enabling fast training and efficient deployment in industrial environments.

Industrial 3D Anomaly Detection

While significant progress has been achieved in 2D anomaly detection, the exploration of anomaly detection in 3D data remains relatively nascent[15]. Recent efforts have begun to bridge this gap by leveraging the rich geometric and structural information available in 3D point clouds and meshes.[49] proposed the Complementary Pseudo Multimodal Feature (CPMF) framework, which combines local geometric features extracted via handcrafted point cloud descriptors with global semantic cues obtained from pseudo-2D projections processed by pretrained 2D networks. Similarly, multimodal approaches such as those by[50] and[51] integrate geometric and photometric features to enhance robustness in defect localization. In contrast, purely geometric methods perform anomaly detection directly in the 3D domain.[16] introduced 3D Student-Teacher (3D-ST), the first unsupervised anomaly detection method to operate exclusively on 3D point clouds. Trained on normal data, it localizes geometric anomalies in high-resolution samples via a single forward pass. 3D-ST extends the student-teacher paradigm to the 3D domain, where a student network is trained to replicate local descriptors generated by a pretrained teacher network. The teacher uses self-supervised learning to extract geometry-aware descriptors by aggregating local features within a controllable receptive field. Anomaly scores are computed based on the deviation between student and teacher predictions. Inspired by PatchCore[13],[17] proposed Reg3D-AD, a registration-based point cloud anomaly detection method that preserves both local and global representations using a dual-feature framework. The method combines raw 3D coordinates with PointMAE features[52], constructing a memory bank of neighborhood-sensitive characteristics derived from normal training data.[18] introduced Group3AD, a group-level feature-based framework for efficient 3D anomaly representation. They propose an Inter-cluster Uniformity Network (IUN) to encourage uniform distribution across feature clusters and an Intra-cluster Alignment Network (IAN) to tightly align features within each cluster. Additionally, an Adaptive Group-Center Selection (AGCS) module is used to enhance localization of potential anomalies based on geometric saliency. Despite promising performance, both Reg3D-AD and Group3AD involve substantial computational and memory overhead.[19] proposed IMRNet, a self-supervised iterative mask reconstruction framework. It employs a geometry-aware sampling module during training to preserve potentially anomalous local structures during downsampling. A transformer reconstructs masked patches in a self-supervised fashion. During inference, the point cloud is processed iteratively, with each reconstructed output becoming the input for the next pass. Anomalies are localized by contrasting the final reconstruction with the original input.[20] introduced R3D-AD, a diffusion-based method for reconstructing anomalous point clouds. This approach leverages the denoising diffusion process to progressively remove aberrant geometries through learned point-wise displacements, ultimately producing a normal version of the input and enabling fine-grained anomaly localization. Despite recent advances, challenges remain in scaling 3D anomaly detection to real-world industrial settings. These include high computational demands, limited labeled datasets, and sensitivity to occlusion and sensor noise. Nonetheless, the integration of 3D reasoning, self-supervision, and generative modeling continues to push the boundaries of precise and automated 3D defect detection.

Motivated by SimpleNet's success[48] as a simple, single-stream CNN for 2D image anomaly detection, we set out to develop an equally efficient architecture for 3D point cloud anomaly detection. However, directly applying SimpleNet[48] to

point clouds is nontrivial: unlike the regular grid structure of images, point clouds are unordered and irregular, requiring permutation-invariant operations and explicit modeling of local geometric relationships; additionally, the higher dimensionality, variable density, and sensor noise inherent in 3D scans demand specialized feature extraction and domain adaptation. Therefore, we propose a novel 3D anomaly detection network that extends SimpleNet's single-pass design by integrating a parameter-free Spatial Context Aggregation module to fuse local neighbourhoods with global prototypes, a lightweight Feature Adaptor for bridging pretraining and industrial domains, and an Anomalous Feature Generator to synthesize hard negatives in feature space. This compact, end-to-end framework preserves SimpleNet's efficiency while addressing the unique challenges of industrial 3D inspection.

Methodology

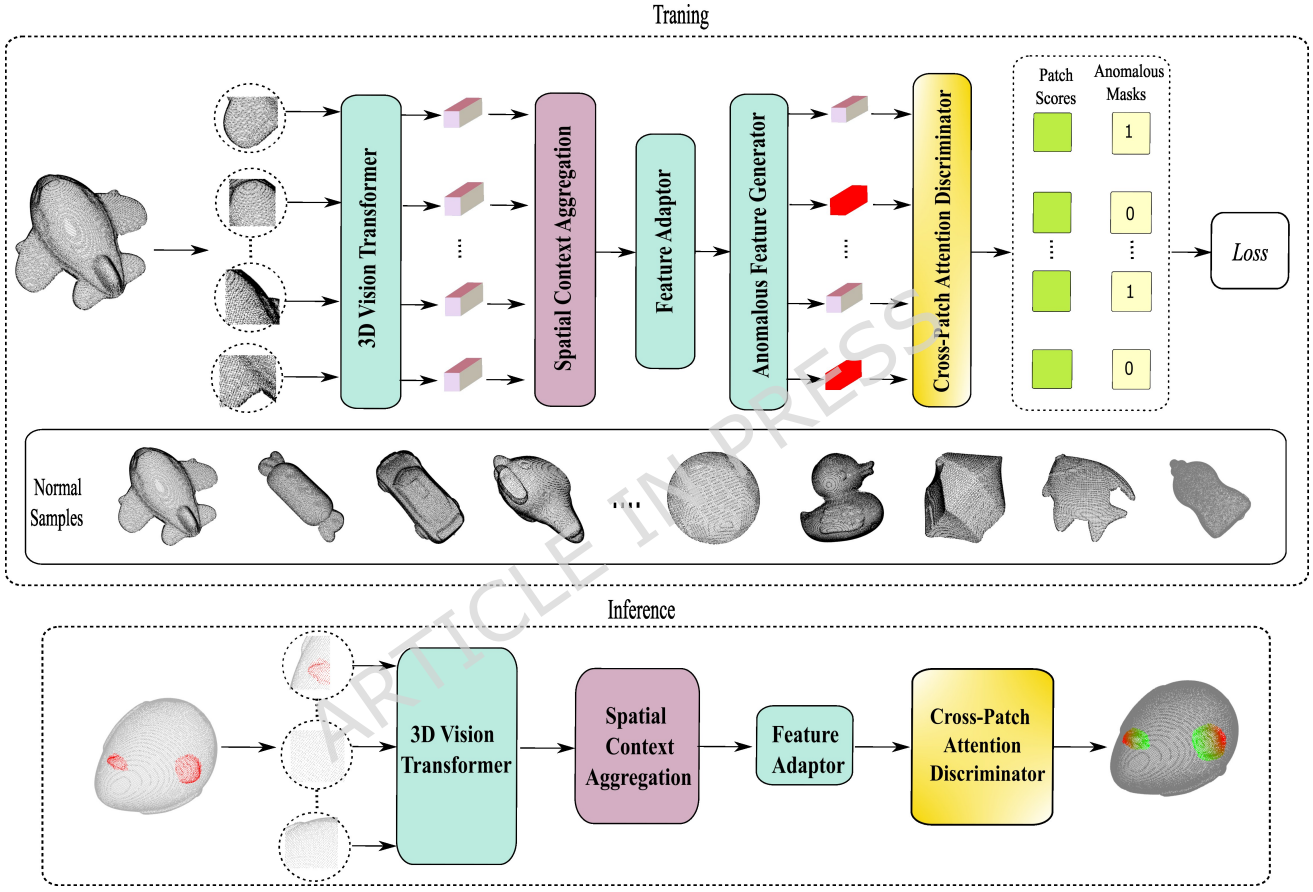

Fig. 1. Overview of the 3D anomaly detection pipeline. The input 3D point cloud is first embedded into patch tokens by a frozen ShapeNet-pretrained Point-MAE backbone. Tokens are then refined by the training-free Spatial Context Aggregation module and aligned to the industrial domain via a lightweight Feature Adaptor. During training only, the Anomalous Feature Generator selectively corrupts a random subset of adapted tokens with scaled Gaussian noise to produce hard negatives. A lightweight cross-patch attention discriminator, trained with a patch-wise binary cross-entropy loss, learns to distinguish clean tokens from corrupted ones. At inference all generator components are disabled and the discriminator directly outputs a dense anomaly score map over the 3D cloud in one pass.

As illustrated in Figure 1, our method formulates 3D point-cloud anomaly detection as a four-stage pipeline. First, overlapping local patches are extracted via farthest-point sampling and k-NN grouping and embedded into C-dimensional tokens using a frozen ShapeNet-pretrained Point-MAE backbone. Second, a parameter-free Spatial Context Aggregation module refines these tokens by fusing local geometric descriptors with global structural prototypes via optimal-transport weighting. Third, a compact two-layer MLP Feature Adaptor bridges the gap between synthetic pretraining and real industrial scans. Finally, during training, a Noise-Random-Patches Anomalous Feature Generator perturbs a subset of tokens with learnably scaled Gaussian noise. A cross-patch attention Discriminator, trained with patch-wise cross-entropy loss, learns to separate clean from corrupted tokens.

At inference, the generator is disabled and the attention discriminator directly produces per-patch anomaly scores that are reprojected onto the original point cloud to yield dense heatmaps and a global anomaly decision in a single forward pass.

Backbone

Our anomaly detection backbone is the Masked Autoencoder for point clouds (Point-MAE[52], which ingests a raw point set

$$X = \{\mathbf{x}_i \in \mathbb{R}^3\}_{i=1}^{N} \tag{1}$$

and first generates n overlapping local patches by farthest-point sampling (FPS) followed by k-nearest-neighbor grouping:

$$C = \mathrm{FPS}(X) \in \mathbb{R}^{n\times 3}, \qquad P_j = \mathrm{KNN}_k\big(X; C_j\big) \in \mathbb{R}^{k\times 3},\ j = 1,\ldots,n. \tag{2}$$

Each patch P_j is normalized by its center C_j and embedded via a lightweight PointNet E_{pt} into a C-dimensional token,

$$T_j = E_{\mathrm{pt}}\big(P_j - C_j\big) \in \mathbb{R}^C, \tag{3}$$

yielding the initial token matrix $T^{(0)} \in \mathbb{R}^{n\times C}$. A stack of L standard Transformer encoder layers then refines these tokens: for $\ell = 1,\ldots,L$,

$$\widetilde{T}^{(\ell)} = \mathrm{SA}\big(\mathrm{LN}(T^{(\ell-1)})\big) + T^{(\ell-1)}, \qquad T^{(\ell)} = \mathrm{MLP}\big(\mathrm{LN}(\widetilde{T}^{(\ell)})\big) + \widetilde{T}^{(\ell)}, \tag{4}$$

where $\mathrm{SA}(Q,K,V) = \mathrm{softmax}(QK^\top/\sqrt{C})V$ is multi-head self-attention and LN denotes layer normalization. The output of Backbone $Z = T^{(L)} \in \mathbb{R}^{n\times C}$ will be used for further stage. Note that we employ the frozen Point-MAE trained on ShapeNet[53].

Spatial Context Aggregation

The designation Spatial Context Aggregation highlights that each patch token is refined by fusing information according to its 3D location (spatial) and both its immediate neighborhood and the broader point cloud structure (context), all via deterministic, parameter-free operations (aggregation). Importantly, this module is entirely *training-free*: it introduces no learnable weights, instead relying on handcrafted geometric descriptors, farthest-point sampling, and optimal-transport-based weight computation[54]. By optimal transport we refer to a principled soft-matching formulation that minimizes a transport cost between two discrete distributions; in practice we compute these soft assignments with an entropic regularized solver whose iterative routine is commonly known as Sinkhorn iterations, which produce stable, approximately doubly-stochastic assignment matrices. This yields a plug-and-play refinement that boosts anomaly sensitivity without risk of overfitting and with minimal computational overhead.

Concretely, let $Z = [z_1,\ldots,z_n]^\top \in \mathbb{R}^{n\times C}$ be the patch-level features and $c_i \in \mathbb{R}^3$ their centers. We first compute for each c_i a descriptor $g_i \in \mathbb{R}^d$ (for example, the Fast Point Feature Histogram (FPFH)[55]), forming $G = [g_1,\ldots,g_n]^\top$. A small subset of $\bar{n} \ll n$ *patch prototypes* $\{p_j\} \subset \{c_i\}$ is chosen by farthest-point sampling. Soft assignments γ_{ij} from patches to prototypes are then obtained by minimizing

$$\sum_{i,j} \gamma_{ij}\Big(\tfrac{1}{\sqrt{D_c}}\|c_i - p_j\|^2 + \tfrac{1}{\sqrt{D_g}}\|g_i - \tilde{g}_j\|^2\Big), \quad \sum_i \gamma_{ij} = \mu_j,\ \sum_j \gamma_{ij} = \tfrac{1}{n}, \tag{5}$$

via Sinkhorn iterations[54], where $\tilde{g}_j$ and $\tilde{z}_j$ denote the prototype's current geometric and feature centroids (updated by $\tilde{g}_j = \sum_i \gamma_{ij} g_i / \sum_i \gamma_{ij}$, $\tilde{z}_j = \sum_i \gamma_{ij} z_i / \sum_i \gamma_{ij}$) and μ_j captures initial density. Each prototype j then gathers its K nearest patches $\mathcal{N}_j$, builds similarity matrices

$$S^g_{j,ii'} = \tfrac{g_i \cdot g_{i'}}{\sqrt{d}}, \quad S^z_{j,ii'} = \tfrac{z_i \cdot z_{i'}}{\sqrt{C}}, \tag{6}$$

which are Sinkhorn-normalized to W^g_j, W^z_j. The two normalized affinity matrices are fused by elementwise product and a softmax

$$W_j \propto \exp\big(W^g_j \odot W^z_j\big), \tag{7}$$

and then used to update the prototype feature

$$\tilde{z}_j \leftarrow \tfrac{1}{2}\Big(\tilde{z}_j + \sum_{i\in\mathcal{N}_j} W_{j,i}\, z_i\Big). \tag{8}$$

After all prototypes are refined, a global aggregation among them proceeds similarly: inter-prototype similarities $\tilde{S}^g_{jk} = (\tilde{g}_j \cdot \tilde{g}_k)/\sqrt{d}$, $\tilde{S}^z_{jk} = (\tilde{z}_j \cdot \tilde{z}_k)/\sqrt{C}$ are computed, distant pairs ($\|p_j - p_k\| > \tau$) are masked, the remaining affinities are normalized and fused into $\tilde{W}$, and then

$$[\tilde{z}_1; \ldots; \tilde{z}_{\bar{n}}] \leftarrow \tfrac{1}{2}\big([\tilde{z}_1; \ldots; \tilde{z}_{\bar{n}}] + \tilde{W}\,[\tilde{z}_1; \ldots; \tilde{z}_{\bar{n}}]\big). \tag{9}$$

Finally, prototype features propagate back to original patches: coordinate weights $w^c_{ij} \propto \exp(-\|c_i - p_j\|^2/\tau_c)$ and feature weights $w^z_{ij} \propto \exp(z_i \cdot \tilde{z}_j/\sqrt{C})$ are normalized by Sinkhorn to yield W, and each

$$z_i \leftarrow \tfrac{1}{2}\Big(z_i + \sum_{j=1}^{\bar{n}} W_{ij}\,\tilde{z}_j\Big). \tag{10}$$

Optionally, a final snapping of z_i to its nearest prototype in cosine space can enforce consistency. With only $\mathcal{O}(n\bar{n} + nK + \bar{n}^2)$ complexity and a handful of hyperparameters (K, τ, τ_c), this Spatial Context Aggregation dramatically enhances backbone features by injecting both local smoothness and global structural coherence, which is crucial for detecting subtle geometric anomalies.

Feature Adaptor

To mitigate the domain gap between the ShapeNet-pretrained Point-MAE and our industrial data, we insert a lightweight non-linear Feature Adaptor G_θ. It is implemented as a two-layer MLP with a bottleneck and normalization:

$$\begin{aligned} h_i &= \mathrm{Norm}_1\big(W^{(1)} z_i + b^{(1)}\big), \qquad W^{(1)} \in \mathbb{R}^{\frac{C}{2} \times C},\ b^{(1)} \in \mathbb{R}^{\frac{C}{2}}, \\ u_i &= \mathrm{LeakyReLU}(h_i), \\ q_i &= \mathrm{Norm}_2\big(W^{(2)} u_i + b^{(2)}\big), \qquad W^{(2)} \in \mathbb{R}^{C \times \frac{C}{2}},\ b^{(2)} \in \mathbb{R}^{C}, \end{aligned} \tag{11}$$

where $z_i \in \mathbb{R}^C$ is the aggregated patch feature from the Spatial Context module, LeakyReLU adds non-linearity, and each Norm_k denotes either Batch-Norm or Layer-Norm (chosen based on validation stability). A residual skip-connection can optionally be added:

$$q_i \leftarrow z_i + \alpha\,(q_i - z_i), \tag{12}$$

with learnable or fixed $\alpha \in [0, 1]$, to preserve original pre-trained representations while allowing flexible adaptation. This MLP adaptor remains lightweight (two linear layers with a single hidden bottleneck) yet has the capacity to reshape feature distributions more richly than a purely linear mapping, stabilizing training across diverse industrial geometries.

Anomalous Feature Generator

Industrial defects in 3D point clouds are typically sparse, subtle, and spatially localized. To simulate such defects during training without requiring external anomaly examples, we introduce a lightweight feature-space perturbation strategy that synthesizes pseudo-anomalous tokens by injecting noise into a randomly selected subset of feature embeddings. This approach is designed to mimic the irregular and partial nature of real-world anomalies while maintaining semantic consistency in the majority of the object.

Given an adapted patch token $q_i \in \mathbb{R}^C$, we first sample a binary mask to determine whether this token should be corrupted. Specifically, we draw:

$$m_i \sim \mathrm{Bernoulli}(p), \tag{13}$$

where $m_i = 1$ indicates corruption and $m_i = 0$ means the token remains clean. The corruption probability $p \in (0, 1)$ controls the sparsity of anomaly injection and is treated as a tunable hyperparameter (typically $p = 0.5$). This Bernoulli masking mechanism reflects the observation that true industrial defects tend to affect only a small fraction of the object's surface, and encourages the model to focus on learning to distinguish localized abnormalities rather than relying on global statistical shifts. For tokens selected for corruption ($m_i = 1$), we generate an additive perturbation vector drawn from a Gaussian distribution:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_C), \tag{14}$$

where σ controls the scale of the noise. Smaller values of σ produce subtle perturbations close to the normal manifold (harder negatives), while larger values result in more distinguishable outliers. We tune σ via grid search on a validation set and find $\sigma = 0.1$ offers a good trade-off across datasets. The final pseudo-anomalous feature is then computed as:

$$\tilde{q}_i = q_i + \gamma m_i \varepsilon_i, \tag{15}$$

where γ is a learnable scalar parameter that adjusts the magnitude of the perturbation during training. This allows the model to adaptively calibrate how strongly the corrupted features should deviate from the clean ones, providing flexibility across different object geometries or training regimes. When $m_i = 0$, the original feature is passed through unchanged, ensuring that the normal distribution remains well-defined and unblurred by unnecessary noise. Importantly, this synthetic anomaly generation mechanism is applied only during training. At inference time, the generator is disabled and all feature tokens $\{q_i\}$ are passed cleanly through the discriminator for scoring, thereby avoiding any stochasticity or overhead during test-time deployment.

Difference to SimpleNet[48]. Our method contrasts with the SimpleNet-style anomaly generator, which adds Gaussian noise indiscriminately to all tokens in the feature space. While this approach is straightforward, it uniformly perturbs the feature manifold and may blur the boundary between normal and anomalous distributions. In contrast, our selective random-patch corruption generates harder and more realistic negatives by targeting only a sparse subset of tokens. This better reflects the distributional asymmetry and local nature of real defects in industrial settings, resulting in improved training signal and more precise anomaly discrimination.

Cross-Patch Attention Discriminator

Anomalies in industrial 3D data often depend not only on local deviations but also on their context within the surrounding geometry. A minor surface defect might be imperceptible in isolation but becomes apparent when contrasted against nearby undisturbed regions. To support such context-aware reasoning, we employ a lightweight attention-based discriminator that considers interactions across all patch tokens jointly, rather than evaluating each token independently.

During training, the discriminator receives both clean tokens $\{q_i\}$ and their selectively corrupted counterparts $\{\tilde{q}_i\}$, generated by the Anomalous Feature Generator. To encode spatial information and preserve the relative positioning of patches, we first add a learnable positional embedding $e_i \in \mathbb{R}^C$ to each token. This yields positionally enriched representations:

$$\hat{q}_i = q_i + e_i, \quad \hat{\tilde{q}}_i = \tilde{q}_i + e_i. \tag{16}$$

The complete set of encoded tokens $\{\hat{q}_i\}$ or $\{\hat{\tilde{q}}_i\}$, depending on whether training or inference is being performed, is then passed through a shallow attention-based network. This consists of a multi-head self-attention layer with four heads (each of size $C/4$) followed by a compact multi-layer perceptron that maps each token to a scalar anomaly logit. Formally, the forward pass computes:

$$\{z_i\} = \mathrm{MHA}(\{\hat{q}_i\}), \tag{17}$$

$$s_i = \mathrm{MLP}(z_i), \tag{18}$$

where MHA denotes multi-head attention, and MLP is a shared feedforward head with one hidden layer. Layer normalization and dropout are included to stabilize training. This architecture allows each token to aggregate information from all others, enabling the model to detect structural inconsistencies and semantic outliers by comparing each region of the object with its global context.

At test time, the Anomalous Feature Generator is disabled and only clean tokens $\{q_i\}$ are evaluated. These are augmented with positional embeddings and passed through the attention-based discriminator to produce per-patch anomaly scores $\{s_i\}$. These scores are reprojected onto the point cloud by assigning each point the maximum score of the patches that contain it. To reduce the impact of isolated false positives and improve spatial smoothness, a 3D median filter is optionally applied to the resulting heatmap.

Difference to SimpleNet. In SimpleNet, anomaly scoring is performed by an independent two-layer multilayer perceptron applied separately to each patch token. This design does not model interactions between tokens and is therefore limited in its ability to detect spatially dependent or globally inconsistent patterns. By contrast, our attention-based discriminator jointly processes all tokens and incorporates positional information, allowing it to learn dependencies across both spatially adjacent and distant regions. This leads to improved sensitivity to subtle and structured defects.

Loss Function and Training

We jointly optimize the Feature Adaptor G_θ and the attention-based Discriminator using a patch-wise binary classification objective. For each patch, the binary label $y_i = m_i$ indicates whether it was corrupted by the Anomalous Feature Generator. The predicted anomaly score is $s_i \in \mathbb{R}$, and the patch-wise binary cross-entropy loss is:

$$\mathscr{L} = -\frac{1}{P}\sum_{i=1}^{P}[y_i \log \sigma(s_i) + (1-y_i)\log(1-\sigma(s_i))] + \lambda \|\theta\|_2^2, \tag{19}$$

where $\sigma(\cdot)$ is the sigmoid function, P is the number of patches per cloud, and λ is the weight decay applied to G_θ. Regularization in the Discriminator is handled by dropout and attention normalization. We use the AdamW optimizer with an initial learning rate of 10^{-4}, cosine annealing over 100 epochs, batch size 8, and gradient clipping with a norm limit of 1. Prior to patch extraction, standard 3D augmentations (random yaw, point jittering, scaling) are applied to increase robustness.

Inference and Scoring Function

At inference, the Anomalous Feature Generator is disabled. Clean patch features $\{q_i\}$ are passed through the trained Discriminator to obtain anomaly logits:

$$s_i = \text{Discriminator}(q_i + e_i). \tag{20}$$

These scores are interpreted as per-patch anomaly intensities. To produce a dense heatmap on the original point cloud, each point is assigned the maximum score from the patches that contain it. A 3D median filter is applied for noise suppression. For segmentation, the heatmap is thresholded at τ_{seg}, selected on a validation set. For global anomaly detection, the point cloud is assigned a score

$$s_{\text{cloud}} = \max_i s_i, \tag{21}$$

and marked anomalous if $s_{\text{cloud}} > \tau_{\text{pc}}$.

Evaluation

Datasets

Real3D-AD is a challenging high-precision dataset tailored for industrial 3D point cloud anomaly detection. It comprises 1,254 point clouds across 12 object categories including airplanes, cars, candy bars, chickens, diamonds, ducks, fish, gemstones, seahorses, shells, starfish, and toffees where each model is represented by between forty thousand and over two million points. Point clouds were acquired using a PMAX-S130 blue-light optical system mounted on a rotating turntable, delivering full 360 degree coverage and drastically reducing ambient light interference. With point-to-point resolutions as fine as 0.0010 m to 0.0015 m, Real3D-AD stands out as the most detailed public dataset for 3D industrial anomaly detection to date. Every category in Real3D-AD includes both pristine prototypes for training and defective samples for testing. Defects are classified into two core types: bulges, representing redundant surface protrusions, and sinks, denoting missing or eroded regions, and are carefully annotated via CloudCompare's octree-based comparison and manual labeling tools. In the dataset structure, each class folder contains a train subfolder with only good point clouds, a test subfolder with defective point clouds, and a ground-truth folder of text files encoding per-point anomaly masks, which enables supervised benchmarking at the point level.

To complement existing high-precision benchmarks and better reflect the constraints of typical factory inspections, we introduce Industrial3D-AD, a novel 3D anomaly detection dataset captured with a widely used industrial depth camera. Our objective was to create a benchmark that not only mimics real-world sensing conditions, complete with ambient lighting variation, quantization noise, and partial occlusions, but also challenges anomaly detectors with defects that are subtle and spatially sparse. Raw depth frames were back-projected into dense 3D point clouds, then downsampled and denoised using a combination of voxel grid filtering and statistical outlier removal. To simulate the variability encountered on production lines, we deliberately adjusted lighting levels and introduced slight misalignments, producing realistic domain shifts across captures. We selected ten industrially relevant artifacts spanning a wide spectrum of geometries, including flat metal plates, cylindrical rods, and intricately cast components, made from diverse materials such as alloys, plastics, ceramics, and painted finishes. For each artifact, we collected 1,000 pristine scans to form a training corpus of 10,000 normal point clouds. Anomaly examples were created by carefully introducing and documenting five defect classes: fine scratches and abrasion networks, small pits and holes (sub-millimeter to a few millimeters in diameter), excess material burrs along edges, localized dents and bends, and

chipped or missing sections. Each anomalous instance was annotated at the point level by projecting the known defect footprints onto the cleaned point clouds and manually refining the labels. The resulting test set comprises 2,000 point clouds balanced between 1,000 defect-free and 1,000 anomalous clouds, each accompanied by a pixel-perfect mask for rigorous evaluation. Industrial3D-AD poses several key challenges: sensor quantization noise that can mimic or obscure true defects, limited spatial resolution that turns sub-millimeter cracks into just a handful of points, and surface reflectivity effects, particularly on glossy or dark finishes, that lead to missing measurements.

Implementation Details

Our pipeline is implemented in Python 3.8 and PyTorch 1.11 with CUDA 11.3 and cuDNN 8.2, and all experiments run on a workstation with an NVIDIA GeForce RTX 4090 GPU. Each raw point cloud is preprocessed by voxel-grid downsampling (voxel size 0.0005 m for Real3D-AD and 0.001 m for Industrial3D-AD), centered to zero mean and scaled to unit radius. During training we apply standard geometric augmentations consisting of random yaw rotation, Gaussian jitter with standard deviation 0.005 m, and uniform scaling in the range plus or minus 5 percent. These augmentation ranges were chosen to reflect typical pose and capture variability observed in our industrial acquisition pipeline and were validated qualitatively and quantitatively on a small hold-out set prior to final experiments.

Patch extraction uses farthest-point sampling to obtain $n = 1024$ patch centers with $k = 32$ nearest neighbors per patch, and each patch is embedded with a lightweight PointNet into a $C = 256$ dimensional token. The choice of n and k trades off spatial coverage and per-patch resolution; we evaluated n in the range 512 to 2048 and k in the range 16 to 64 and observed diminishing returns in point-level AUPR beyond $n = 1024$ and $k = 32$, while computational cost grew linearly. The backbone is a frozen Point-MAE Transformer with $L = 12$ encoder layers (8 heads, hidden MLP size $4C$) pretrained on ShapeNet. Freezing the backbone preserves pretrained geometric priors while enabling fast adaptation.

Spatial Context Aggregation selects $\bar{n} = 0.1n = 100$ prototypes by farthest point sampling on the patch centers. This prototype fraction was selected by sweeping $\bar{n}/n$ between 0.05 and 0.25 and choosing the smallest value that reaches a performance plateau on validation AUPR while keeping compute affordable. For each prototype we gather $K = 10$ neighboring patches; K is set relative to the patch size so that prototype aggregation captures local neighborhood geometry without excessive smoothing. Prototype to patch affinities are computed using a fused geometry and feature similarity and are normalized via Sinkhorn optimal transport. We set the Sinkhorn temperature to $t = 0.05$ and run $T_{\mathrm{sk}} = 20$ iterations in our default configuration, and we use the same (t, T_{sk}) pair for the multiple Sinkhorn normalization steps within SCA. Specifically, we observed that with temperature in the range 0.02 to 0.08 the transport plan stabilizes and that 10 to 30 iterations produce near-identical assignment matrices up to numerical precision while $T_{\mathrm{sk}} = 20$ offers a robust compromise between assignment fidelity and wall-clock time on our hardware.

Several SCA hyperparameters control locality and fusion. The mask radius τ and coordinate bandwidth τ_c were chosen by a small grid search, with τ controlling the spatial extent for local fusion and τ_c scaling the coordinate kernel used in prototype assignment. The reported defaults $\tau = 0.5$ and $\tau_c = 0.1$ represent settings that consistently produced strong object-level consistency across both datasets. The Feature Adaptor is a deliberately compact two layer MLP with hidden size $C/2 = 128$, LeakyReLU slope 0.1, and a residual interpolation weight $\alpha = 0.5$. In our final runs both normalization layers Norm_1 and Norm_2 are implemented as Layer Normalization to improve stability at the chosen batch size. This architecture was selected to limit overfitting while still enabling effective re-centering and scaling of pretrained tokens. We compared adaptor capacities ranging from a single linear layer up to three-layer MLPs. Performance saturates at the two-layer design on our validation folds, and larger adaptors increased sensitivity to noise in small validation sets.

The Anomalous Feature Generator perturbs a randomly selected subset of patch tokens with probability $p = 0.5$ using isotropic Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ scaled by a learnable factor γ. The choice of p and σ was determined by grid search over $p \in \{0.3, 0.5, 0.7\}$ and $\sigma \in \{0.05, 0.1, 0.2\}$ using validation AUPR as the selection criterion. We observed that too small a corruption probability produces trivially easy negatives while too large a probability erodes the normal manifold; $p = 0.5$ balanced hardness and realism. The default noise level $\sigma = 0.1$ achieved stable discrimination of localized anomalies without destabilizing training.

The Discriminator comprises a multi-head self-attention block with 4 heads followed by a compact MLP head with one hidden layer (hidden size $4C$, dropout 0.1). Positional embeddings $e_i \in \mathbb{R}^C$ are learnable, initialized from $\mathcal{N}(0, 0.02^2)$, and trained jointly with the Feature Adaptor and the Discriminator. We train only the Feature Adaptor and the Discriminator. Optimization uses AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1 \times 10^{-8}$, initial learning rate 1×10^{-4} decayed with a cosine schedule over 100 epochs, weight decay 1×10^{-4}, gradient clipping at norm 1, and batch size 8 per GPU. The L2 regularization term $\lambda \|\theta\|_2^2$ in the loss is implemented via the optimizer weight decay with $\lambda = 1 \times 10^{-4}$. These solver settings follow common practice for small adapter networks, and we validated the learning rate and weight decay by running short runs on a validation fold prior to full training.

Validation set construction and hyperparameter selection follow a reproducible protocol. For each dataset we create a

validation split by stratified sampling over object classes and defect categories so that the validation set reflects the same class and defect-type distribution as the training set. When defect-type labels are scarce we instead ensure coverage by sampling at the object instance level and by enforcing that each class contributes a minimum number of examples to validation. Hyperparameters were tuned by maximizing point-level AUPR on this validation split, which better reflects precision-recall trade-offs in highly imbalanced anomaly detection tasks than AUROC alone. For segmentation threshold selection we perform a separate grid search on the validation fold and select the threshold τ_{seg} that maximizes the F1 score at the object aggregation level, which we found to align well with operational needs for downstream inspection. At inference we disable anomaly synthesis and run a single forward pass through Backbone, Spatial Context Aggregation, Feature Adaptor, and Discriminator. Patch scores are projected to points using a max pooling over containing patches, a 3D median filter of radius three voxels is applied to suppress isolated noise, and the segmentation threshold τ_{seg} is applied.

Evaluation Metrics

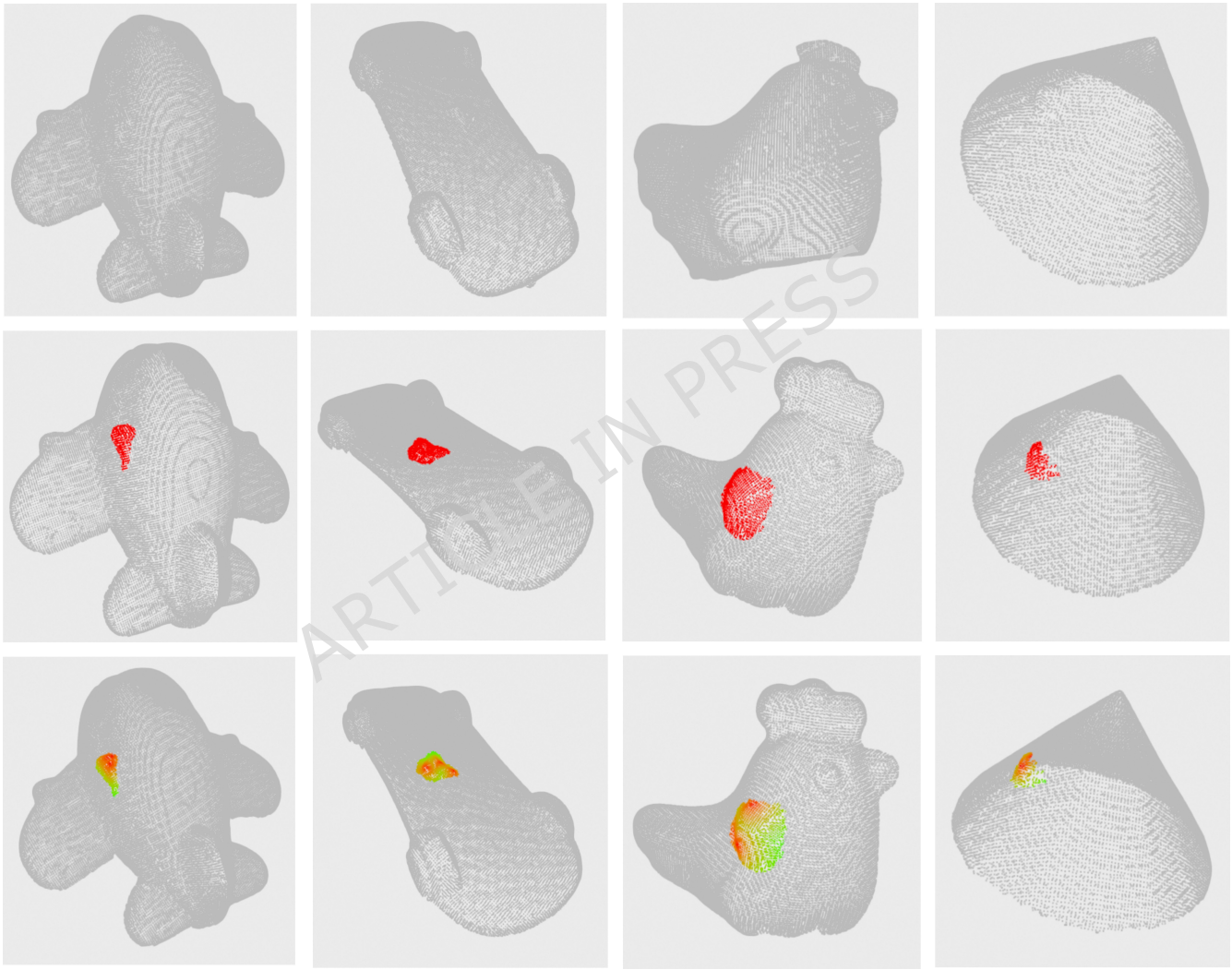

Fig. 2. Qualitative results on Real3D-AD dataset. From top to bottom: input point clouds, Ground truth annotations of anomalous points in red, and anomaly scores for each 3D point predicted by the proposed method.

To evaluate anomaly detection methods on Real3D-AD, we employ both object-level and point-level metrics derived from the receiver operating characteristic (ROC) curve and the precision-recall (PR) curve. The ROC curve illustrates the trade-off between sensitivity and specificity by plotting the true positive rate (TPR) against the false positive rate (FPR) as the decision threshold varies:

$$\text{True Positive Rate (TPR)} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \qquad \text{False Positive Rate (FPR)} = \frac{\text{FP}}{\text{FP}+\text{TN}}.$$

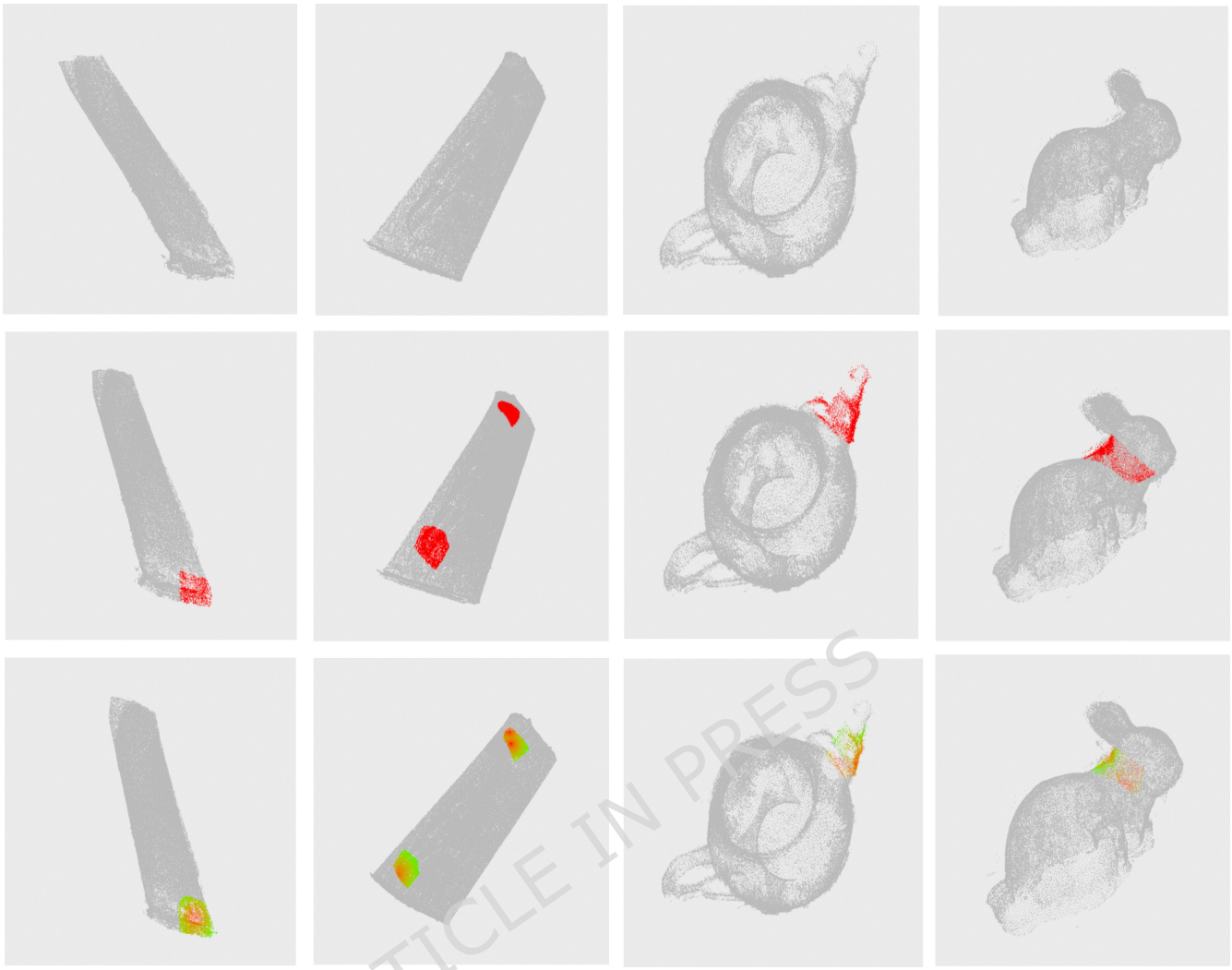

Fig. 3. Qualitative results on the newly collected Industrial3D-AD dataset. From top to bottom: input point clouds, Ground truth annotations of anomalous points in red, and anomaly scores for each 3D point predicted by the proposed method.

The area under the receiver operating characteristic (AUROC) integrates this curve into a single scalar measure:

$$\text{Area Under the Receiver Operating Characteristic (AUROC)} = \int_0^1 \text{TPR}\big(\text{FPR}\big)\,\text{dFPR},$$

where an AUROC of 0.5 corresponds to random classification and 1.0 indicates perfect discrimination. The AUROC is threshold-independent and robust to class imbalance, making it a widely used performance metric. The precision-recall (PR) curve focuses on the positive (anomalous) class by plotting precision against recall (which equals TPR) over all thresholds. Precision is defined as

$$\text{Precision } (P) = \frac{\text{TP}}{\text{TP}+\text{FP}},$$

and the area under the precision-recall curve (AUPR) is given by

$$\text{Area Under the Precision}-\text{Recall Curve (AUPR)} = \int_0^1 P(R)\,\text{d}R.$$

Because the baseline precision equals the fraction of anomalies in the dataset, AUPR is particularly informative in highly imbalanced settings, reflecting a model's ability to retrieve true anomalies without generating excessive false alarms. At the object level, each point cloud is treated as a single instance-typically summarized by its maximum per-point anomaly score, and

AUROC and AUPR are computed across all point clouds. At the point level, each point's anomaly score is compared directly against its ground-truth mask label, yielding fine-grained localization performance.

Result

Table 1. Average Results on Real3D-AD.

	Point Level		Object Level		Speed
	AUROC	AUPR	AUROC	AUPR	FPS
BTF(Raw)	0.571	0.022	0.603	0.611	2.05
BTF(FPFH)	0.730	0.064	0.635	0.614	1.01
M3DM(PointMAE)	0.637	0.046	0.552	0.572	0.30
M3DM(PointBERT)	0.636	0.052	0.538	0.581	0.50
PatchCore(FPFH)	0.577	0.071	0.593	0.591	0.09
PatchCore(FPFH+Raw)	0.680	0.123	0.682	0.667	0.09
PatchCore(PointMAE)	0.642	0.058	0.594	0.633	0.10
3D-ST[16]	0.705	0.109	0.645	0.723	1.50
Reg3D-AD[17]	0.705	0.109	0.704	0.723	0.08
Group3AD[18]	0.735	0.137	0.751	0.740	2.50
IMRNet[19]	0.725	0.166	0.725	0.625	5.60
R3D-AD[20]	0.592	0.041	0.734	0.632	7.12
Ours	0.763	0.194	0.781	0.775	13.52

Figures 2 and 3 present qualitative segmentation results on the Real3D-AD and Industrial3D-AD benchmarks, respectively, illustrating the model's ability to localize both bulges and sinks in high-precision scans and to robustly highlight subtle scratches, dents, and occlusion-induced artifacts under factory-like capture conditions. Tables 1 and 2 present a comparison between the proposed method and state-of-the-art approaches. For transparency in runtime comparison, all Speed (FPS) entries were measured on an NVIDIA GeForce RTX 4090. Where available, we ran each method using the authors' public implementation on our workstation and report the resulting end-to-end inference throughput under the same preprocessing and evaluation protocol to ensure a consistent comparison. BTF(Raw) refers to the use of only the raw 3D coordinate features within the Back-to-Front framework[15]. BTF(FPFH) augments the same pipeline with Fast Point Feature Histograms[55]. M3DM(PointMAE) and M3DM(PointBERT) correspond to the multimodal architecture configured to ignore its RGB branch and instead extract point cloud features with PointMAE[52] or PointBERT[56]. PatchCore variants replace the usual ResNet-based feature extractor with either hand-crafted descriptors or pretrained point backbones before applying PatchCore scoring.

Table 1 reveals clear causal links between the proposed architectural components and the observed performance improvements. At the point level, our method achieves an AUROC of 0.763 and an AUPR of 0.194, surpassing the previous best, Group3AD[18], by +2.8% in AUROC and +5.7% in AUPR. The larger relative gain in AUPR reflects improved precision-recall balance in this highly imbalanced point-level detection task and arises primarily from two interacting design choices. First, the Feature Adaptor mitigates covariate mismatch between the frozen, shape-pretrained backbone and the industrial scan domain. This alignment enhances the model's sensitivity to subtle defect signatures, increasing true positive detections while keeping false positives low. Second, the Selective Anomalous Feature Generator introduces sparse, hard negative tokens during training, which sharpens decision boundaries in feature space and suppresses spurious activations on clean surfaces. Together, these modules strengthen the model's ability to isolate weak, localized anomalies from noisy background structures, explaining why the improvement in AUPR is more pronounced than that in AUROC.

At the object level the model reaches 0.781 AUROC and 0.775 AUPR, which reflects improved spatial consistency of anomalous regions across patches. This improvement is closely tied to the Spatial Context Aggregation module. By aligning patch tokens with geometry-aware prototypes and propagating affinities via an optimal transport based assignment, the module enforces local smoothness while preserving global structure. This mechanism reduces isolated false positives that would otherwise disrupt object-level aggregation and it increases the likelihood that contiguous anomalous regions produce consistently elevated scores across multiple neighboring patches. In contrast, approaches that rely on per-patch independent scoring or on large memory banks may either miss subtle, distributed defects or produce unstable object-level signals. Transformer-based and distillation-based baselines such as M3DM variants and 3D-ST yield competitive localization accuracy, but in our experiments they do not surpass the balanced precision-recall trade-off achieved by combining adaptor alignment, selective feature corruption, and prototype-guided context aggregation, as reflected by their lower AUPR values.

The per-method breakdown also sheds light on why some baselines perform differently across metrics. Hand-crafted descriptors such as FPFH can deliver strong AUROC when geometric discontinuities are pronounced, but they are less expressive

for fine-grained surface texture and therefore show limited AUPR. Memory-bank approaches such as PatchCore attain high precision in some regimes but suffer from heavy computational costs and sensitivity to the choice of stored exemplars, leading to lower operational throughput and sometimes brittle recall. Heavy transformer encoders that are trained or distilled on 2D-3D multimodal data provide rich context but become computationally expensive and less effective when RGB cues are removed or when the pretraining domain diverges from industrial geometries.

Table 2. Performance comparison on the newly collected Industrial3D-AD dataset in terms of point-level and object-level AUROC, AUPR, and inference speed (FPS).

	Point Level		Object Level		Speed
	AUROC	AUPR	AUROC	AUPR	FPS
BTF(Raw)	0.545	0.021	0.577	0.585	2.05
BTF(FPFH)	0.702	0.061	0.610	0.586	1.01
M3DM(PointMAE)	0.612	0.044	0.528	0.545	0.30
M3DM(PointBERT)	0.608	0.049	0.512	0.553	0.50
PatchCore(FPFH)	0.548	0.068	0.563	0.561	0.09
PatchCore(FPFH+Raw)	0.648	0.117	0.650	0.637	0.09
PatchCore(PointMAE)	0.615	0.055	0.567	0.604	0.10
3D-ST[16]	0.670	0.104	0.614	0.689	1.50
Reg3D-AD[17]	0.678	0.104	0.670	0.690	0.08
Group3AD[18]	0.701	0.131	0.716	0.708	2.50
IMRNet[19]	0.691	0.158	0.691	0.597	5.60
R3D-AD[20]	0.563	0.039	0.699	0.603	7.12
Ours	0.730	0.184	0.744	0.741	13.52

Table 2 shows results on our newly collected Industrial3D-AD dataset, which features a wider range of component geometries, finer defect scales, and material surface variations compared to Real3D-AD. At the point level our method achieves 0.730 AUROC and 0.184 AUPR, outperforming Group3AD by +2.9% in AUROC and +5.3% in AUPR. The overall lower absolute scores relative to Real3D-AD reflect the added difficulty of shiny surfaces, subtle defect signatures, and irregular sensor noise patterns. From a causal perspective these domain factors diminish the signal-to-noise ratio of pointwise cues and increase within-class variance, which in turn reduces both precision and recall for small anomalies. In such cases the Feature Adaptor and Spatial Context Aggregation still provide tangible benefits by stabilizing feature statistics and encouraging spatial coherence, but their effectiveness is limited where geometric descriptors become unstable due to extreme sparsity or specular noise.

The relative improvements on Industrial3D-AD provide additional insight into module contributions. The maintained advantage in AUPR indicates that selective generation and discriminator training continue to suppress false positives on complex surfaces, while prototype-guided propagation helps aggregate weak per-point signals into coherent object-level detections. At the same time the residual performance gap versus Real3D-AD highlights structured failure modes where the current isotropic feature corruptions do not adequately simulate extended defects such as long scratches or delaminations. This observation motivates future extensions that combine geometry-level synthesis and targeted sensor simulations to cover a broader spectrum of defect morphologies.

In terms of inference speed, the pipeline runs at 13.52 FPS, significantly faster than most baselines. This efficiency is a direct consequence of the single-pass design, the frozen backbone, and the compact Feature Adaptor and discriminator. The Spatial Context Aggregation performs prototype alignment and propagation using a lightweight optimal transport solver and does not require large exemplar stores or nearest neighbor retrieval at test time, which keeps runtime low. Memory-bank and reconstruction methods incur substantial overhead from exemplar matching or multi-pass reconstruction and therefore trade off throughput for certain types of precision. Our results show that the proposed combination of accuracy and speed is achievable by carefully balancing pretrained representation power with small, task-specific modules that impose geometric consistency and robust decision boundaries.

Overall, the quantitative and qualitative results support the view that the measured gains arise from complementary mechanisms. The Feature Adaptor improves alignment and recall, the Selective Anomalous Feature Generator increases robustness to hard negatives and raises precision, and the Spatial Context Aggregation imposes spatial consistency that benefits object-level aggregation and reduces isolated false alarms. These causal links explain why our method attains stronger AUPR and object-level metrics while remaining computationally efficient. We discuss remaining limitations and concrete directions for extending these mechanisms in the Discussion section.

Ablation Study

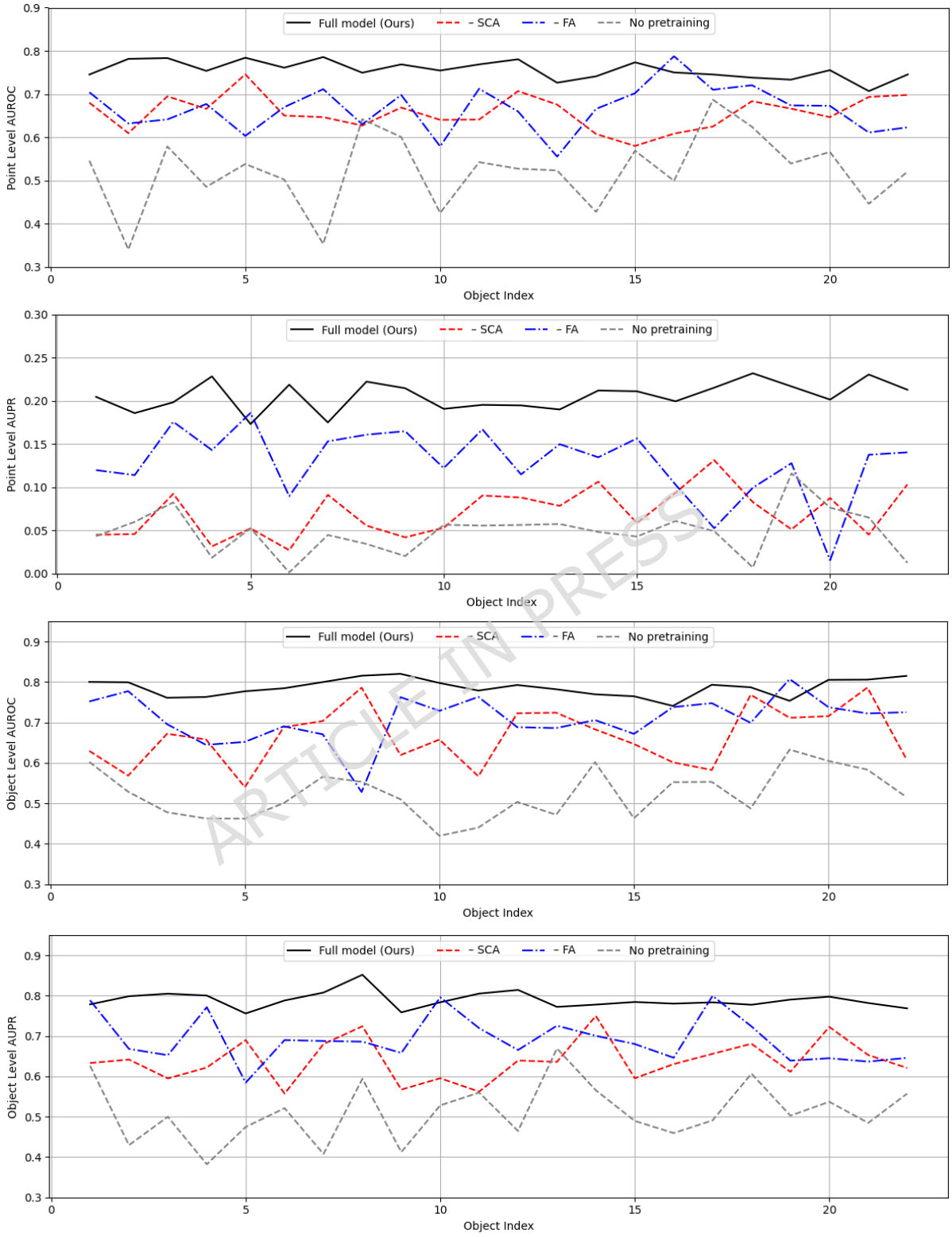

Fig. 4. Object- and point-level performance comparison across 22 test objects. Four line plots show (top left) point-level AUROC, (top right) point-level AUPR, (bottom left) object-level AUROC, and (bottom right) object-level AUPR for the full model (black solid) and three ablated variants (- SCA in red dashed; - FA in blue dash-dot; no pretraining in gray dashed).

Table 3. Ablation Study: Average Results Across Datasets (AFG always present).

	Point Level		Object Level		Speed
	AUROC	AUPR	AUROC	AUPR	FPS
Full model (Ours)	0.750	0.181	0.770	0.775	13.52
- SCA	0.719	0.103	0.715	0.707	18.66
- FA	0.735	0.169	0.752	0.756	14.79
AFG (SimpleNet)	0.722	0.143	0.731	0.734	13.52
Discriminator (SimpleNet)	0.724	0.132	0.725	0.729	14.15
No global propagation	0.725	0.143	0.733	0.734	16.52
No prototype snapping	0.743	0.165	0.755	0.760	14.22
Pure geometry (no W^z)	0.731	0.149	0.734	0.736	14.46
Pure feature (no W^g)	0.733	0.145	0.736	0.732	14.72
No pretraining	0.504	0.053	0.525	0.530	13.52

To quantify the impact of every design choice and hyperparameter in our pipeline, we conduct all ablations under the same training setups and baseline settings (frozen Point-MAE backbone; Spatial Context Aggregation with $\bar{n} = 0.1n$, $K = 10$, $\tau = 0.5$, $\tau_c = 0.1$; Feature Adaptor MLP hidden size $C/2$, residual weight $\alpha = 0.5$; Anomalous Feature Generator corruption probability $p = 0.5$ and noise standard deviation $\sigma = 0.1$; 3D median filter radius 3 voxels at inference). We evaluate: (i) module removals: – SCA (omit Spatial Context Aggregation) and – FA (bypass the Feature Adaptor); (ii) SCA internals: No global propagation (skip inter-prototype fusion), No prototype snapping (disable final cosine re-assignment), Pure geometry (fuse only W^g) and Pure feature (fuse only W^z); (iii) learning setup: No pretraining (train transformer end-to-end); and (iv) core module baselines: AFG (SimpleNet), which perturbs all tokens with Gaussian noise rather than selectively, and Discriminator (SimpleNet), which replaces the attention-based discriminator with a per-token two-layer MLP.

Figure 4 illustrates that our full model consistently delivers the best and most stable point- and object-level AUROC/AUPR across all 22 test objects. Table 3 reports the impact of each variant on average point- and object-level AUROC/AUPR as well as inference speed.

The full model achieves the highest performance across all metrics, attaining a point-level AUROC of 0.750 and AUPR of 0.181, along with object-level scores of 0.770 AUROC and 0.775 AUPR, while operating at 13.52 frames per second (FPS). Among all variants, removing the Spatial Context Aggregation (SCA) module leads to the most substantial performance degradation. Specifically, point-level AUROC and AUPR drop to 0.719 and 0.103, respectively, and object-level AUROC and AUPR fall to 0.715 and 0.707. Although inference speed improves to 18.66 FPS without SCA, the significant accuracy loss underscores the essential role of local-global fusion in enhancing the expressiveness of geometric features and capturing subtle surface anomalies.

Eliminating the Feature Adaptor (FA) results in a more moderate decline in detection quality. In this configuration, point-level AUROC and AUPR decrease to 0.735 and 0.169, while object-level AUROC and AUPR reduce to 0.752 and 0.756. The slight improvement in inference speed to 14.79 FPS suggests that the adaptor introduces only a marginal computational cost. Nevertheless, these results indicate that aligning the pretrained backbone features with the target industrial domain enhances anomaly separability and provides measurable gains in precision, even if it is not strictly indispensable.

We also compare our selective Anomalous Feature Generator with a variant inspired by SimpleNet, which perturbs all patch tokens uniformly by adding Gaussian noise. This modification leads to a clear performance drop: point-level AUPR declines from 0.181 to 0.143 and object-level AUPR from 0.775 to 0.734. These results validate the advantage of our selective corruption strategy, where only a random subset of tokens is perturbed. By introducing sparse and localized feature distortions, the generator produces harder negative samples that more accurately simulate realistic defects and improve the training signal for the discriminator.

Replacing our attention-based Discriminator with a SimpleNet-style per-token two-layer MLP leads to further degradation in performance. The point-level AUPR drops to 0.132 and object-level AUPR to 0.729, with a modest increase in speed to 15.85 FPS. This decline highlights the importance of joint reasoning across patch tokens: cross-patch attention enables the model to capture spatial dependencies and contextual relationships between patches, which are particularly important for identifying semantic inconsistencies and subtle geometric anomalies. In contrast, treating each token independently limits the model's ability to distinguish structural outliers that manifest only in relation to their neighbors.

Further ablations explore the internal mechanisms of the Spatial Context Aggregation module. Disabling global propagation between prototypes results in notable performance loss, with AUPR reduced to 0.143 at the point level and 0.734 at the object level, even though speed increases to 16.52 FPS. This confirms that long-range context transfer among prototypes contributes meaningfully to the model's ability to reason about distant but structurally related regions. Disabling the final prototype

snapping or restricting fusion to only geometric (W^g) or only learned feature (W^z) affinities yields milder declines in detection quality. For all these variants, point-level AUROC remains above 0.731 and AUPR above 0.145, while object-level scores also stay above 0.732. These findings indicate that both affinity modalities are beneficial and that the soft re-assignment step, though not critical alone, adds refinement that enhances overall robustness.

Finally, removing the ShapeNet pretraining and training the Point-MAE backbone from scratch causes detection performance to collapse: point-level AUROC and AUPR fall to 0.504 and 0.053, and object-level scores drop to 0.525 and 0.530, while runtime remains unchanged. We attribute this dramatic degradation to the loss of broad geometric priors that large-scale self-supervised pretraining encodes. ShapeNet pretraining provides stable, multi-scale shape representations that make patch tokens semantically meaningful and that in turn allow the Spatial Context Aggregation and prototype assignments to form coherent groups. When the backbone is learned from scratch under the weak, synthetic supervision regime of our discriminator, its early-layer features remain noisy and unstable, the adaptor lacks the capacity and data to fully recover these priors, and the discriminator is forced to separate poorly formed tokens which encourages overfitting to spurious cues. The severe class imbalance of pointwise anomaly detection further amplifies this effect, since few reliable positive patterns are available to guide a large randomly initialized transformer.

This ablation study provides strong empirical support for our architectural choices. The Spatial Context Aggregation module and pretrained backbone are essential for accurate and reliable anomaly detection. The Feature Adaptor and global prototype interactions offer measurable improvements at minimal computational cost. Most notably, our selective pseudo-anomaly generation and attention-based discrimination outperform the corresponding SimpleNet variants by a clear margin, validating the core innovations of our method in terms of both accuracy and robustness in industrial 3D anomaly detection scenarios.

Discussion

Robustness to domain variation is a central concern for industrial anomaly detection, and it deserves careful consideration together with the architectural contributions presented in this work. The proposed pipeline was designed with two complementary inductive biases that improve transferability across sensing conditions. The first is the lightweight Feature Adaptor, which explicitly recenters and rescales pretrained backbone tokens to match the statistical properties of the target scanner. The second is the Spatial Context Aggregation module, which aligns learned patch tokens with geometry-aware prototypes and propagates contextual information through optimal transport based affinity. These two components together mitigate moderate shifts in sensor noise, viewpoint, and surface reflectivity, and they form the main mechanisms by which the model adapts to the statistics of unseen scans. Nevertheless, these components do not remove all sources of domain shift, and the method retains certain limitations that also suggest clear directions for improvement.

Differences in sensing physics and measurement noise present the first major challenge. Distinct devices and acquisition parameters can introduce heteroskedastic noise, anisotropic sampling, and specular artifacts that are not fully represented by the augmentations or feature corruptions used during training. The selective anomalous feature generator encourages the discriminator to detect local deviations from the normal feature manifold, but it is not a complete model of all physical sensor effects. When reflectivity or multipath interference produces structured outliers or dense spurious regions, performance may degrade unless representative data or sensor-specific simulations are incorporated. Practical mitigations include calibration, a small amount of in-domain clean data for adaptor tuning, and the introduction of physically informed noise models during pretraining or adaptation. Extending the current corruption strategy to explicitly model such sensor-dependent effects is a promising future direction.

Point density and sampling heterogeneity also affect robustness. The fixed-radius and fixed-k nearest neighbor patch extraction used in this study assumes a relatively stable range of densities. When point clouds are either extremely sparse or strongly nonuniform, the local geometric descriptors used for prototype selection can become unstable, and the receptive field of each patch may vary across the surface. Multi-scale patching, adaptive neighborhood selection, and density-normalized feature aggregation are natural extensions that can alleviate this sensitivity. A hierarchical tokenization scheme in which coarse tokens encode global context and fine tokens capture local geometry would further improve resilience to variations in sampling resolution and object complexity.

Another limitation arises from the diversity of real-world defect types. The selective generator synthesizes sparse, hard negatives that emulate many small and localized anomalies such as dents or small missing regions. However, extended or structured defects, including long scratches or large missing parts, may not be fully captured by isotropic perturbations in feature space. Future extensions may incorporate geometry-level corruption, spatially coherent deformations, or physics-based simulations of material defects. Coupling such structured anomaly synthesis with limited in-domain fine-tuning or self-supervised adaptation could substantially enhance coverage of diverse defect modalities.

Finally, certain hyperparameters of the Spatial Context Aggregation step, including the number of prototypes, regularization strength, and the number of Sinkhorn iterations, introduce algorithmic choices that may require modest adjustment across domains. Although these parameters are not learned, they trade off computational efficiency with assignment fidelity. For

deployment under resource constraints, prototype sub-sampling or approximate assignment could be explored. Likewise, while the Feature Adaptor is intentionally lightweight to prevent overfitting, larger or dynamically parameterized adaptors could be evaluated in future work when greater domain mismatch is expected.

Conclusion

We presented a practical, single-pass framework for unsupervised 3D anomaly detection that is specifically designed for industrial inspection. The method combines a parameter-free Spatial Context Aggregation module, a compact two-layer Feature Adaptor for domain alignment, and a Selective Anomalous Feature Generator that synthesizes hard negatives in feature space. Together with a lightweight attention-based discriminator and a frozen pretrained Masked Autoencoder backbone, these components deliver a balance of geometric sensitivity and computational efficiency without relying on memory banks or multi-pass inference. Empirically, the proposed pipeline attains state-of-the-art detection performance on both Real3D-AD and our Industrial3D-AD benchmark, with point- and object-level gains of up to 5.7% in AUPR and 3.0% in AUROC over prior strong baselines. The implementation runs at 13.52 FPS on an NVIDIA GeForce RTX 4090 under a unified evaluation protocol, demonstrating that the approach is lightweight and suitable for near-real-time inspection in practical manufacturing settings. In future work, we will pursue engineering and algorithmic optimizations to further improve throughput and edge deployment readiness. Promising directions include model pruning and quantization, domain-tailored self-supervised pretraining for different sensor modalities, and integration with robotic inspection pipelines to enable automated corrective actions.

Data Availability

The newly collected Industrial3D-AD dataset used and analyzed during the current study are available from the corresponding author on reasonable request. The Real3D-AD dataset is accessible on GitHub at https://github.com/M-3LAB/Real3D-AD.

References

1. Pu, C. *et al.* Geometric spatial constraints network for slender and tiny surface defect detection. *Adv. Eng. Informatics* **65**, 103138 (2025).
2. Ma, X. *et al.* Stamping part surface crack detection based on machine vision. *Measurement* **251**, 117168 (2025).
3. Hoang, D.-C. *et al.* Unsupervised industrial anomaly detection using paired well-lit and low-light images. *J. Comput. Des. Eng.* **12**, 41–61 (2025).
4. Zhang, K. *et al.* Automatic measurement system for aircraft rivet flushness on surfaces empowered by multi-modal large-scale models. *Adv. Eng. Informatics* **69**, 103936 (2026).
5. Egodawela, S. *et al.* Metal loss defect detection and depth estimation using multi-spectral image analysis of cooling excited steel specimen with corrosion. *Sci. Reports* **15**, 23894 (2025).
6. Hoang, D.-C. *et al.* Unsupervised visual-to-geometric feature reconstruction for vision-based industrial anomaly detection. *IEEE Access* (2025).
7. Tong, X. *et al.* Dam-faster rcnn: few-shot defect detection method for wood based on dual attention mechanism. *Sci. Reports* **15**, 22860 (2025).
8. Hoang, D.-C. *et al.* Image-based anomaly detection in low-light industrial environments with feature enhancement. *Results Eng.* **25**, 104309 (2025).
9. Zhang, C., Guo, Z. & Li, C. Unsupervised anomaly detection for gearboxes based on the deep convolutional support generative adversarial network. *Sci. Reports* **15**, 20977 (2025).
10. Song, H. Rstd-yolov7: a steel surface defect detection based on improved yolov7. *Sci. Reports* **15**, 19649 (2025).
11. Bergmann, P., Fauser, M., Sattlegger, D. & Steger, C. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4183–4192 (2020).
12. Shi, Y., Yang, J. & Qi, Z. DFR: Deep feature reconstruction for unsupervised anomaly segmentation. *Neurocomputing* **424**, 9–22 (2021).
13. Roth, K. *et al.* Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14318–14328 (2022).

14. Hoang, D.-C. *et al.* Accurate industrial anomaly detection with efficient multimodal fusion. *Array* 100512 (2025).

15. Horwitz, E. & Hoshen, Y. 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*, 2968–2977 (2023).

16. Bergmann, P. & Sattlegger, D. Anomaly detection in 3d point clouds using deep geometric descriptors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2613–2623 (2023).

17. Liu, J. *et al.* Real3D-AD: A dataset of point cloud anomaly detection. *Adv. Neural Inf. Process. Syst.* **36**, 30402–30415 (2023).

18. Zhu, H. *et al.* Towards high-resolution 3d anomaly detection via group-level feature contrastive learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4680–4689 (2024).

19. Li, W. *et al.* Towards scalable 3d anomaly detection and localization: A benchmark via 3d anomaly synthesis and a self-supervised learning network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22207–22216 (2024).

20. Zhou, Z. *et al.* R3D-AD: Reconstruction via diffusion for 3d anomaly detection. In *European Conference on Computer Vision*, 91–107 (Springer, 2024).

21. Venkataramanan, S., Peng, K.-C., Singh, R. V. & Mahalanobis, A. Attention guided anomaly localization in images. In *European Conference on Computer Vision*, 485–503 (Springer, 2020).

22. Ding, C., Pang, G. & Shen, C. Catching both gray and black swans: Open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7388–7398 (2022).

23. Chadha, G. S., Rabbani, A. & Schwung, A. Comparison of semi-supervised deep neural networks for anomaly detection in industrial processes. In *2019 IEEE 17th international conference on industrial informatics (INDIN)*, vol. 1, 214–219 (IEEE, 2019).

24. Chu, W.-H. & Kitani, K. M. Neural batch sampling with reinforcement learning for semi-supervised anomaly detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, 751–766 (Springer, 2020).

25. Gu, Z. *et al.* Remembering normality: Memory-guided knowledge distillation for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16401–16409 (2023).

26. Tong, G., Li, Q. & Song, Y. Enhanced multi-scale features mutual mapping fusion based on reverse knowledge distillation for industrial anomaly detection and localization. *IEEE Transactions on Big Data* **10**, 498–513 (2024).

27. Hou, J. *et al.* Divide-and-Assemble: Learning block-wise memory for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8791–8800 (2021).

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

29. Zavrtanik, V., Kristan, M. & Skočaj, D. DSR–a dual subspace re-projection network for surface anomaly detection. In *European conference on computer vision*, 539–554 (Springer, 2022).

30. Fang, Z. *et al.* FastRecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17481–17490 (2023).

31. Zhang, X., Xu, M. & Zhou, X. RealNet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16699–16708 (2024).

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

33. Song, J., Kong, K., Park, Y.-I., Kim, S.-G. & Kang, S.-J. AnoSeg: Anomaly segmentation network using self-supervised learning. *arXiv preprint arXiv:2110.03396* (2021).

34. Liang, Y. *et al.* Omni-frequency channel-selection representations for unsupervised anomaly detection. *IEEE Transactions on Image Process.* **32**, 4327–4340 (2023).

35. De Nardin, A., Mishra, P., Foresti, G. L. & Piciarelli, C. Masked transformer for image anomaly localization. *Int. J. Neural Syst.* **32**, 2250030 (2022).

36. Yao, X., Li, R., Qian, Z., Luo, Y. & Zhang, C. Focus the discrepancy: Intra-and inter-correlation learning for image anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6803–6813 (2023).

37. Luo, W., Yao, H., Yu, W. & Li, Z. AMI-Net: Adaptive mask inpainting network for industrial anomaly detection and localization. *IEEE Transactions on Autom. Sci. Eng.* (2024).

38. Zhang, X. *et al.* Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6782–6791 (2023).

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

40. Tebbe, J. & Tayyub, J. Dynamic addition of noise in a diffusion model for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3940–3949 (2024).

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

42. Wan, Q., Gao, L., Li, X. & Wen, L. Unsupervised image anomaly detection and segmentation based on pretrained feature mapping. *IEEE Transactions on Ind. Informatics* **19**, 2330–2339 (2022).

43. Defard, T., Setkov, A., Loesch, A. & Audigier, R. PaDiM: a patch distribution modeling framework for anomaly detection and localization. In *International conference on pattern recognition*, 475–489 (Springer, 2021).

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

45. Zuo, Z., Wu, Z., Chen, B. & Zhong, X. A reconstruction-based feature adaptation for anomaly detection with self-supervised multi-scale aggregation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5840–5844 (IEEE, 2024).

46. Hyun, J. *et al.* ReConPatch: Contrastive patch representation learning for industrial anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2052–2061 (2024).

47. Li, C.-L., Sohn, K., Yoon, J. & Pfister, T. CutPaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9664–9674 (2021).

48. Liu, Z., Zhou, Y., Xu, Y. & Wang, Z. SimpleNet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20402–20411 (2023).

49. Cao, Y., Xu, X. & Shen, W. Complementary pseudo multimodal feature for point cloud anomaly detection. *Pattern Recognit.* **156**, 110761 (2024).

50. Wang, Y. *et al.* Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8032–8041 (2023).

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

52. Pang, Y. *et al.* Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*, 604–621 (Springer, 2022).

53. Chang, A. X. *et al.* Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

54. Peyré, G., Cuturi, M. *et al.* Computational optimal transport: With applications to data science. *Foundations Trends Mach. Learn.* **11**, 355–607 (2019).

55. Rusu, R. B., Blodow, N. & Beetz, M. Fast point feature histograms (FPFH) for 3d registration. In *2009 IEEE international conference on robotics and automation*, 3212–3217 (IEEE, 2009).

56. Yu, X. *et al.* Point-BERT: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19313–19322 (2022).

Author Contributions

All authors contributed equally to the conceptualization, formal analysis, investigation, methodology, and writing and editing of the original draft. All authors have read and agreed to the published version of the manuscript.

Competing Interests

The authors declare no competing interests.

Funding Declaration

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.