Information Fusion 130 (2026) 104097

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

IAENet: An importance-aware ensemble model for 3D point cloud-based anomaly detection

Xuanming Cao [a], Chengyu Tao [b], Yifeng Cheng [a], Juan Du [a,b,c,*]

[a] Smart Manufacturing Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, 511453, China
[b] Interdisciplinary Programs Office, The Hong Kong University of Science and Technology, Hong Kong SAR, 999077, China
[c] Department of Mechanical and Aerospace Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, 999077, China

ARTICLE INFO

Keywords:
3D point cloud
Anomaly detection
Ensemble model
Quality inspection
Pretrained representation

ABSTRACT

Surface anomaly detection is pivotal for ensuring product quality in industrial manufacturing. While 2D image-based methods have achieved remarkable success, 3D point cloud-based detection remains underexplored despite its richer geometric cues. We argue that the key bottleneck is the absence of powerful pretrained foundation backbones in 3D comparable to those in 2D. To bridge this gap, we propose Importance-Aware Ensemble Network (IAENet), an ensemble framework that synergizes 2D pretrained expert with 3D expert models. However, naively fusing predictions from disparate sources is non-trivial: existing strategies can be affected by a poorly performing modality and thus degrade overall accuracy. To address this challenge, We introduce a novel Importance-Aware Fusion (IAF) module that dynamically assesses the contribution of each source and reweights their anomaly scores. Furthermore, we devise critical loss functions that explicitly guide the optimization of IAF, enabling it to combine the collective knowledge of the source experts but also preserve their unique strengths, thereby enhancing the overall performance of anomaly detection. Extensive experiments show that IAENet achieves a new state-of-the-art for point-level localization and ranks second at object-level on MVTec 3D-AD dataset. On the Eyecandies dataset, it achieves the best performance in both levels. Additionally, it substantially reduces false positive rates, underscoring its practical value for industrial deployment.

1. Introduction

Surface anomaly detection plays an essential role in ensuring that products meet standards and specifications across a multitude of industrial applications [1]. Traditionally, this task has been heavily dependent on manual visual inspection [2], which is not only labor-intensive but also highly subjective and inefficient. With the advancement of machine vision technology, the possibility of replacing manual inspection with automated methods has become increasingly feasible, offering higher efficiency, accuracy, and robustness [3].

Although well-established 2D image-based detection methods have yielded promising results [4,5], they are not without limitations. These methods are often influenced by lighting conditions and may struggle to identify geometric anomalies that are similar in color to the background. Additionally, they lack the capacity to discern geometric details such as size and shape [1,6]. Due to the rapid advancement in sensing technology [7,8], the acquisition of high-quality 3D point cloud data has become more accessible and affordable, leading to a surge in research focused on 3D point cloud-based anomaly detection [9].

Due to the scarcity of anomalous samples in industrial settings, most anomaly detection methods are unsupervised, focusing on learning the distribution of normal samples. For instance, Roth et al. [4] introduce a memory bank approach where test sample features are compared against stored normal features to identify anomalies. Another method [10] employs a teacher-student architecture to generate anomaly scores based on feature discrepancies between a pretrained teacher model and a student model trained on normal data. In addition to these unimodal methods, recent studies have explored the fusion of image and point cloud data. A representative example [11] uses Fast Point Feature Histograms (FPFH) [12] descriptors and 2D pretrained models to extract multimodal features and construct a memory bank for anomaly detection. Further related work will be detailed in Section 2.

Given the current focus on multimodal data fusion in research, we have observed that anomaly detection based solely on point cloud data is underexplored. However, in many scenarios, point cloud data does not have corresponding RGB images available. Unfortunately, current approaches do not fully exploit the abundant geometric information contained within 3D point clouds. The primary problem is the lack of

* Corresponding author.
E-mail addresses: xcao743@connect.hkust-gz.edu.cn (X. Cao), ctaoaa@connect.ust.hk (C. Tao), yifengcheng@hkust-gz.edu.cn (Y. Cheng), juandu@ust.hk (J. Du).

https://doi.org/10.1016/j.inffus.2025.104097
Received 31 July 2025; Received in revised form 12 November 2025; Accepted 22 December 2025
Available online 25 December 2025

pretrained foundation models in the 3D domain, akin to those in the 2D domain, which possess powerful feature representation capabilities. This gap motivates us to adapt 2D pretrained foundation models by projecting 3D point clouds into textured 2D depth images, and integrate them with 3D expert models to construct an ensemble framework, termed as Importance-Aware Ensemble Network (IAENet), thereby enhancing 3D anomaly detection. As shown in Fig. 1, when fed with textured depth maps, the 2D expert is capable of effectively extracting rich semantic information from fine details, tending to identify all potential anomalies, which results in higher scores for normal points. However, the 3D expert captures more global information and struggles to detect subtle anomalies such as tiny holes in cable gland. Our IAENet model, on the other hand, effectively integrates the predictions from both experts, yielding more accurate detection results.

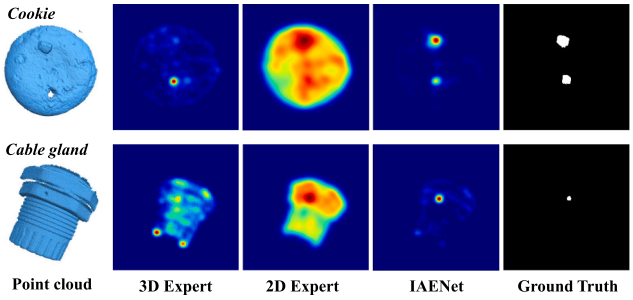

Fig. 1. The illustration of the anomaly score maps of different expert models on two representative objects from MVTec 3D-AD [7]. Despite receiving the same point cloud input, the 2D and 3D experts exhibit complementary strengths. When an anomaly is detected by one expert but not by another, our IAENet integrates both insights to accurately identify all anomalies. It also effectively suppresses the anomaly scores of normal points, thereby reducing the false positive rate.

However, effectively integrating anomaly detection predictions from different source models remains a challenge. Although numerous multimodal studies have explored feature-level fusion, decision-level fusion, which involves the combination of predictions from different modalities, is often simplistic through basic strategies such as addition [11] or taking the maximum value [13] for anomaly detection. These straightforward fusion tactics can be significantly undermined when there is a substantial discrepancy between the predictions of different modalities. This can lead to the fusion decision being adversely affected by the poorer-performing modality, resulting in a combined performance that is inferior to that of the individual modalities operating independently.

The main limitation of existing methods is that they assume equal contributions from the predictions of different models to the final results of anomaly detection. In reality, different models exhibit varying strengths across different scenarios (as demonstrated in Fig. 1), meaning that existing methods cannot adaptively determine the importance under different conditions. To address this, inspired by the research on feature selection [14] which focuses on identifying important features, this paper proposes a novel Importance-Aware Fusion (IAF) module. The IAF module evaluates the contributions of anomaly scores from different models to the final anomaly detection, leading to more effective and robust anomaly detection results.

In general, we propose IAENet, a unified ensemble model that incorporates an innovative decision-level fusion module known as IAF for 3D point cloud-based anomaly detection. Specifically, the model begins with a phase of expert learning, where it trains a set of complementary 2D and 3D expert models to produce per-point predictions. These predictions are derived through feature extraction and a dual memory-bank retrieval process. A lightweight selector network then evaluates the contributions of these predictions to the anomaly detection task and assigns importance scores accordingly. These scores are used to weight the original predictions from the source models, which are then fed into a predictor network for nonlinear mapping to produce the final decision. Our IAENet not only integrates the collective knowledge of the source models but also preserves their unique strengths, thereby enhancing the overall performance of 3D anomaly detection.

To sum up, the contributions of this paper are as follows:

- We propose IAENet, a novel ensemble framework that combines a 2D pretrained foundation model with a dedicated 3D expert to improve both accuracy and robustness in point cloud-based anomaly detection.
- We introduce an innovative Importance-Aware Fusion (IAF) module capable of adaptively evaluating the contributions of predictions from different models to the anomaly detection task. This module aims to preserve the unique strengths of individual models while combining the collective knowledge of the source experts.
- We design critical loss functions to ensure that the IAF module can accurately estimate the contributions of different models and obtain correct predictive results. This is essential for the IAF module to learn the appropriate importance scores and effectively integrate the information from various sources.
- Our IAENet achieves state-of-the-art results on the MVTec 3D-AD dataset, demonstrating a lower false positive rate, which is particularly valuable in industrial applications.

This paper is organized as follows: Section 2 provides a comprehensive review of the related literature, Section 3 details the methodology of our IAENet, Section 4 presents the experimental setup and results, and Section 5 concludes the paper with a summary of our findings and contributions.

2. Related work

In the related work section, we provide a comprehensive review that covers 2D anomaly detection, 3D anomaly detection, and score map fusion strategies.

2.1. 2D anomaly detection

Anomaly detection (AD) based on 2D images has been extensively researched [15,16], with the majority of studies validating their approaches on the MVTec-AD dataset [17]. In general, these methods can be broadly categorized into three main types: normalizing flow-based, knowledge distillation-based and memory bank-based methods. We will briefly introduce them as follows.

Normalizing flow-based methods model complex normal data distributions and flag low-likelihood samples as anomalies [18,19]. Knowledge distillation-based approaches transfer a pretrained teacher to a student fine-tuned on normal data and detect anomalies via teacher-student discrepancies [20–22]. Memory bank-based methods store prototypical normal features for nearest-neighbor comparison at inference and form the backbone of many recent 3D AD systems due to their robustness and adaptability [4,11,13,23,24]. Finally, pretrained image foundation models (e.g., ResNet [25], ViT [26]) have been leveraged to provide strong features for anomaly detection [27,28].

While numerous 2D anomaly detection methods have been developed, they cannot be directly applied to 3D detection problems due to the inherent differences in data representation and complexity [2].

2.2. 3D anomaly detection

Unsupervised anomaly detection based on 3D point clouds was under-researched until the introduction of the MVTec 3D-AD [7], Eyecandies [29], and Real3D-AD [30] datasets. Our discussion here will focus on their 3D point cloud-based branches only.

3D feature extraction. Early 3D anomaly detection relies on handcrafted descriptors like FPFH, which capture local geometry and generalize well. Deep learning advances point-cloud modeling: PointNet

[31] handles unstructured points via point-based multilayer perceptron (MLP) and permutation-invariant aggregation, while transformer-based encoders such as PointMAE [32] model long-range dependencies and are often used as pretrained backbones. Another successful paradigm renders 3D data to 2D and leverages powerful image models: CPMF [33] uses multi-view depth renderings with a pretrained ResNet, and PointCLIP [34] integrates text for multimodal understanding. In this work we adopt the render-to-2D strategy, combining pretrained ResNet features with 3D representations to benefit from both domains.

Feature-based AD. Feature-based methods model the distribution of normal features and detect deviations at test time. 3D-ST [10] uses a teacher-student scheme with RandLA-Net [35] as the point encoder to handle dense scans. AST [21] introduces an asymmetric teacher-student network to further improve detection performance. BTF [11] combines local 3D descriptors (e.g., FPFH) with a memory bank to represent normal geometry. CPMF [33] renders multiple depth maps, extracts 2D pretrained features, and builds a memory bank for detection. M3DM [24] employs a pretrained PointMAE as the 3D encoder. Shape-Guided [13] models local geometry via neural implicit signed distance functions (SDFs) and defines an SDF-based anomaly metric, achieving state-of-the-art results for 3D anomaly localization.

Reconstruction-based AD. Reconstruction-based methods such as IMRNet [36] and 3DSR [37] detect anomalies from reconstruction errors [38], offering high-resolution delineation of anomaly boundaries but sometimes producing false positives on complex surfaces [2]. Given that high-quality point cloud reconstruction remains a challenge, feature-based methods continue to dominate 3D anomaly detection.

Zero-shot and untrained AD. These approaches do not require any training data, including normal and anomalous samples. Zero-shot methods that utilize Vision Language Models (VLM) have recently emerged: PointAD [39] adapts CLIP [40] through prompt learning for zero-shot 3D anomaly detection. Untrained methods [1,41] do not rely on pretrained models; instead, they model anomalies based on the intrinsic features of surfaces, but domain knowledge should be known in advance for modeling.

Our method belongs to the feature-based family and requires training on anomaly-free data. Through the proposed IAF module we effectively fuse complementary information from 2D and 3D experts, improving sensitivity to subtle anomalies while suppressing spurious scores in normal regions. Compared to reconstruction-based approaches, this leads to fewer false positives. Detailed experimental results are reported in the experimental section.

2.3. Score map fusion

In the realm of multimodal methods, the strategies employed for score map fusion are relatively simplistic. The Shape-Guided [13] utilizes a maximum strategy, in which the final fused score map is composed by taking the maximum pixel-wise values from the score maps derived from the SDF and RGB modalities. BTF [11] concatenates features from RGB and point cloud modalities and then performs detection based on a memory bank, which is equivalent to a linear score fusion [23]. M3DM [24], on the other hand, adopts a data-driven method for fusion, employing a one-class Support Vector Machine (OCSVM) to achieve this integration. However, these methods fail to effectively evaluate the contributions of different modalities to the final anomaly detection outcome. Consequently, the fusion results can be significantly impacted by the poorer-performing modality.

Even when using a 3D pretrained model like PointMAE in M3DM, point-level localization may still lag behind classic descriptors such as FPFH; this may stem from smaller pretraining corpora for 3D (e.g., ShapeNet [42]) versus large-scale image datasets like ImageNet [43]. To address these limitations, we propose an ensemble that combines 2D pretrained foundation models with a strong 3D expert–PointNet augmented by SDF-based modeling [13]–and introduce a decision-level fusion module (IAF) that preserves each expert's strengths. Unlike cross-modal alignment methods [24,40] that map different modalities into a shared feature space, IAF operates directly on pixel/point-level anomaly score maps: it preserves modality-specific semantics and uncertainty, is more robust to weak or missing modalities, and adaptively downweights spurious signals while preserving complementary cues. This design mitigates the negative impact of weaker modalities and improves overall detection performance.

3. Methodology

3.1. Overview

Given a point cloud $\mathbf{P} \in \mathbb{R}^{M\times 3}$ representing an arbitrary geometric surface, where M denotes the number of points, the objective of point-level anomaly detection or localization is to localize the anomaly positions, i.e., to output anomaly scores $\mathbf{A} \in \mathbb{R}^{M}$ for all points. Here, anomalous points are assigned higher scores compared to normal points. Additionally, object-level anomaly detection aims to output a single anomaly score $s \in \mathbb{R}$ for the entire point cloud. In this paper, we focus primarily on the unsupervised anomaly detection task, which is more aligned with practical manufacturing scenarios. Specifically, we have access to only an anomaly-free point cloud dataset $\mathcal{D} = \{\mathbf{P}\}$ for training. During the testing phase, the model is expected to directly predict the point-level anomaly scores $\mathbf{A}$ and the object-level anomaly score s for testing point cloud data.

As illustrated in Fig. 2, our IAENet comprises two steps: (i) *source expert learning* and (ii) *importance-aware fusion (IAF)*. The first step trains complementary 2D and 3D experts on $\mathcal{D}$, producing per-point predictions via feature extraction and dual memory-bank retrieval (detailed in Section 3.2). The second step feeds these predictions into a *selector network* that quantifies each expert's reliability for every point, producing importance scores. A *predictor network* then leverages these weights to fuse the expert outputs, preserving their unique strengths while suppressing weaknesses.

To optimize IAF, we synthesize an augmented dataset $\mathcal{D}' = \{(\mathbf{P}'_i, \mathbf{Y}_i)\}_{i=1}^{N}$ by injecting controlled anomalies into $\mathcal{D}$ using a Cut-Paste strategy [23]. Here, $\mathbf{Y}_i$ represents the labels corresponding to $\mathbf{P}'_i$, ranging from 0 to c, where $c = 1$, indicating the presence or absence of anomalies. This supervision allows to learn robust integration without ever seeing real anomalies during training.

3.2. Source expert learning

3.2.1. 3D expert $\mathcal{E}_{3d}$

For the 3D expert, we adopt a PointNet-SDF pipeline [13]. PointNet [31] acts as a patch-wise feature extractor that is first pretrained on the anomaly-free set $\mathcal{D}$ by implicit surface reconstruction. Each patch of the input cloud $\mathbf{P}$ is mapped to a latent vector $\mathbf{f}_1 \in \mathbb{R}^{d_1}$. Given a query point $\mathbf{q}$, the latent code and the query are concatenated and fed into a lightweight decoder ψ that predicts the signed distance

$$s = \psi(\mathbf{q}, \mathbf{f}_1). \tag{1}$$

During pretraining, the decoder is optimized to drive s to zero on the manifold, thus learning an accurate normal surface representation. At anomaly detection step, the 3D expert operates with frozen parameters, and the magnitude $|s|$ serves as the anomaly score for each point, with larger values indicating stronger deviation from the learned normal geometry.

3.2.2. 2D expert $\mathcal{E}_{2d}$

To effectively leverage the powerful feature representation capabilities of 2D foundation models, we adopt a frozen ResNet encoder as our 2D expert. Each point cloud is first rendered into a depth map via normal-based depth rendering [33], which faithfully preserves fine-grained geometric textures, offering richer information than simple depth projection methods. This textured depth map is then processed by the pretrained ResNet to extract global semantic features. Finally,

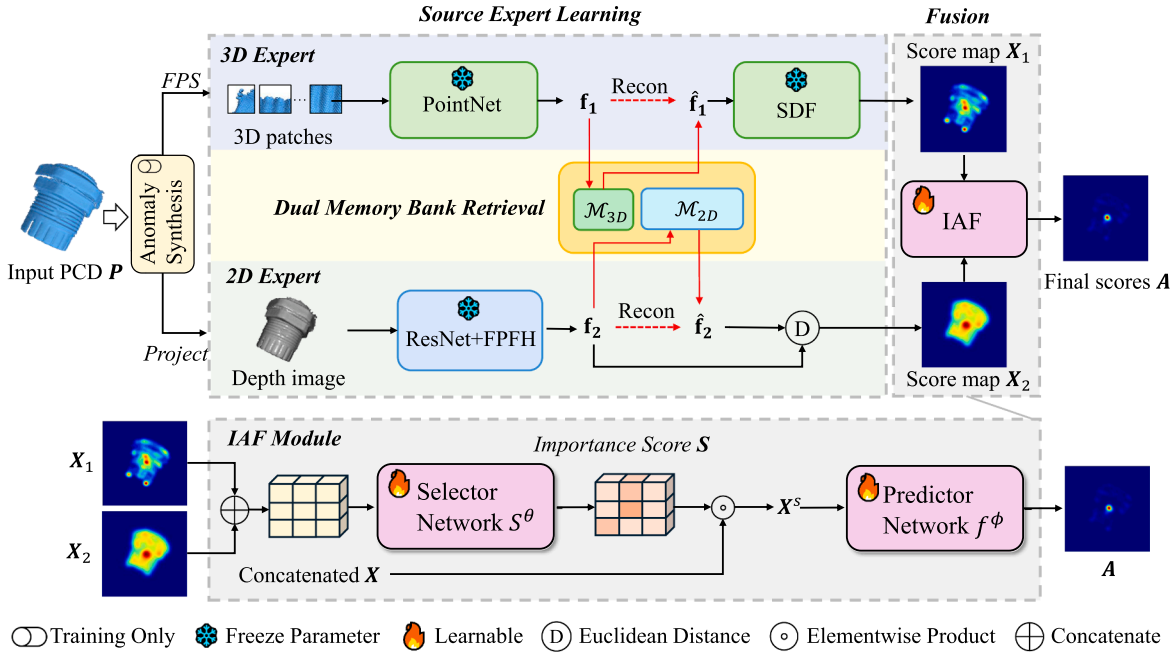

Fig. 2. The framework of our IAENet. During training, normal samples first pass through an anomaly synthesis module that generates pseudo anomalies (disabled at inference). The resulting anomalous point clouds are then processed by two source experts, yielding anomaly score maps $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. Finally, the Importance-Aware Fusion (IAF) module adaptively reweights these scores and outputs the refined anomaly map $\mathbf{A}$, effectively capitalizing on the complementary merits of both experts while suppressing their individual weaknesses.

the global representation is concatenated with per-point FPFH descriptors to yield the composite feature $\mathbf{f}_2 \in \mathbb{R}^{d_2}$ that encodes both global context and local geometry.

3.2.3. Dual memory bank retrieval

Following PatchCore [4], IAENet first populates two dedicated memory banks ($\mathcal{M}_{3D}, \mathcal{M}_{2D}$) with features extracted from the anomaly-free set D. Instead of PatchCore's coreset subsampling, we adopt shape-guided retention [13] to preserve semantically salient features.

Population. For every 3D patch feature $\mathbf{f}_1$ retained in $\mathcal{M}_{3D}$, we store the corresponding 2D features $\{\mathbf{f}_2\}$ of all points within its receptive field into $\mathcal{M}_{2D}$, as illustrated in Fig. 3. This cross-modal pairing guarantees that fine-grained geometry ($\mathbf{f}_1$) and global contextual cues ($\mathbf{f}_2$) are jointly indexed.

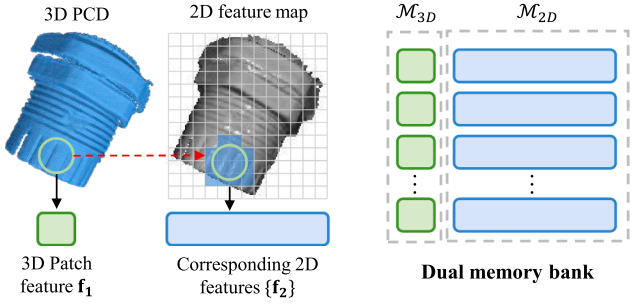

Fig. 3. Illustration of the construction of shape-guided dual memory bank.

Inference. Given a test sample, we process it through both the 3D and 2D branches as follows:

- *3D branch:* patch feature $\mathbf{f}_1$ retrieves k_1 nearest neighbors from $\mathcal{M}_{3D}$ to yield a reconstructed $\hat{\mathbf{f}}_1$. A query point $\mathbf{q}$ in the patch neighborhood is then processed by Eq. (1), producing anomaly map $\mathbf{X}_1$.

$$\mathbf{X}_1(\mathbf{q}) = \left|\psi\left(\mathbf{q}, \hat{\mathbf{f}}_1\right)\right|. \tag{2}$$

- *2D branch:* for each point $\mathbf{p}$, its 2D feature $\mathbf{f}_2(\mathbf{p})$ searches k_2 nearest neighbors in $\mathcal{M}_{2D}$ to obtain the reconstructed feature $\hat{\mathbf{f}}_2(\mathbf{p})$. The anomaly score is

$$\mathbf{X}_2(\mathbf{p}) = \|\mathbf{f}_2(\mathbf{p}) - \hat{\mathbf{f}}_2(\mathbf{p})\|_2. \tag{3}$$

This dual retrieval strategy exploits complementary 2D and 3D cues while keeping the banks compact and interpretable.

For the object-level anomaly score, the maximum values of the score maps are taken, i.e.,

$$s_1 = \max(\mathbf{X}_1), \quad s_2 = \max(\mathbf{X}_2). \tag{4}$$

Here, s_1 and s_2 represent the object-level anomaly scores derived from the score maps $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively.

3.3. Importance-aware fusion (IAF)

Given the prior anomaly maps $\mathbf{X}_1 \in \mathbb{R}^{h\times w}$ and $\mathbf{X}_2 \in \mathbb{R}^{h\times w}$ obtained from source expert learning, the objective is to effectively combine these maps to achieve superior detection results. However, existing methods often fail to accurately assess the reliability of anomaly detection results from different experts. When there is a significant disparity between the two maps, the final result can be severely skewed by the poorer-performing map.

This limitation motivates the development of our IAF module $\mathcal{F}$. IAF learns an adaptive weighting that quantifies the trustworthiness of every spatial location in $\mathbf{X}_1$ and $\mathbf{X}_2$. These learned importance scores are then used to reweight and merge the two maps, allowing the ensemble to leverage their complementary strengths while protecting the output from deficiencies of any individual expert.

3.3.1. Architecture of $\mathcal{F}$

The IAF model $\mathcal{F}$ consists of two main components: the selector network S^θ and the predictor network f^ϕ. The selector network S^θ is constructed using a set of shared MLPs [31], with the specific architecture depicted in Fig. 4. Its primary function is to evaluate the contributions of the two input score maps $\mathbf{X}_1$ and $\mathbf{X}_2$ (both resized into

vectors $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{hw\times 1}$) to the final anomaly detection task and to produce corresponding importance scores $\mathbf{S} \in \mathbb{R}^{hw\times 2}$ that act as weights. This process is formulated as

$$\mathbf{S} = S^{\theta}\left(\mathbf{X}_1 \oplus \mathbf{X}_2\right) = S^{\theta}(\mathbf{X}), \tag{5}$$

where $\oplus$ is the concatenation.

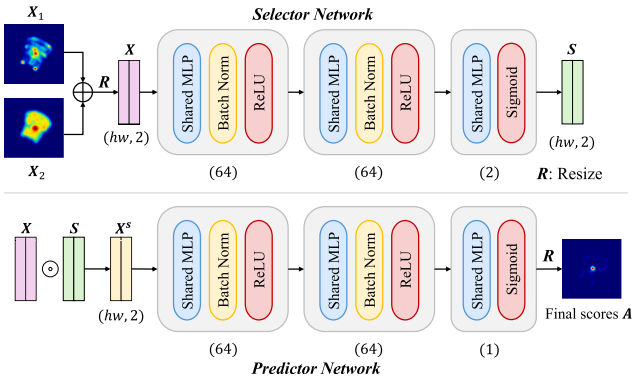

Fig. 4. The architecture of proposed selector network and predictor network.

The predictor network f^{ϕ} then takes the weighted score map as input and performs a nonlinear transformation to obtain the final fused anomaly score. Specifically, the weighted score map $\mathbf{X}^{\mathbf{S}}$ is computed by multiplying each input score map $\mathbf{X}$ with its corresponding importance score $\mathbf{S}$ generated by the selector network. This weighted score map $\mathbf{X}^{\mathbf{S}}$ is then fed into the predictor network, which consists of multiple nonlinear layers designed to capture complex patterns and interactions within the data, as shown in Fig. 4. The output of the predictor network is the final anomaly score $\mathbf{A}$.

$$\mathbf{A} = f^{\phi}(\mathbf{X}, \mathbf{S}) = f^{\phi}(\mathbf{X} \odot \mathbf{S}) = f^{\phi}\left(\mathbf{X}^{\mathbf{S}}\right), \tag{6}$$

where $\odot$ is elementwise product.

3.3.2. Loss functions

Problem formulation. The objective of this study is to develop an importance-aware fusion module $\mathcal{F}$ that outperforms individual experts $\mathcal{E}_{2d}$ and $\mathcal{E}_{3d}$ in anomaly detection task. This is formalized as

$$\mathcal{P}(\mathcal{F}, \mathbf{Y}) > \mathcal{P}\left(\mathcal{E}_{2d}, \mathbf{Y}\right), \mathcal{P}\left(\mathcal{E}_{3d}, \mathbf{Y}\right), \tag{7}$$

where $\mathcal{P}$ denotes the performance of the model relative to target labels $\mathbf{Y}$. Performance is quantified using cross-entropy (CE) [44], leading to the constraint

$$\mathrm{CE}(\mathcal{F}, \mathbf{Y}) < \mathrm{CE}\left(\mathcal{E}_{2d}, \mathbf{Y}\right), \mathrm{CE}\left(\mathcal{E}_{3d}, \mathbf{Y}\right), \tag{8}$$

or equivalently,

$$C_{\mathcal{F}} < b = \min(C_{2d}, C_{3d}). \tag{9}$$

Here, $C_{\mathcal{F}}$ is the cross-entropy loss of the fused model, C_{2d} and C_{3d} are the baseline cross-entropy losses, and b is the minimum of these two baselines. Note that b is a constant, as the input score maps do not change during training.

Selector loss. We reformulate the objective function (9) as an optimizable loss function:

$$\mathcal{R} = \max\left(m - \left(b - C_{\mathcal{F}}\right), 0\right), \tag{10}$$

where m denotes the margin parameter, and $C_{\mathcal{F}}$ is defined as

$$C_{\mathcal{F}} = -\mathbb{E}_{\mathbf{X}\sim P}\left[\sum_{i=1}^{c} y_i \log\left(p\left(y_i|\mathbf{X}^{\mathbf{S}}\right)\right)\right], \tag{11}$$

where $\mathbf{X}$ is the random variable with density P, and the probability $p(y_i|\mathbf{X}^{\mathbf{S}})$ is given by the predictor network f^{ϕ} as follows

$$p\left(y_i|\mathbf{X}^{\mathbf{S}}\right) = f_i^{\phi}(\mathbf{X}, \mathbf{S}). \tag{12}$$

Then, the stochastic form of $\mathcal{R}$ is

$$r(\mathbf{X}, \mathbf{S}) = \max\left(m - \left(b + \sum_{i=1}^{c} y_i \log\left(f_i^{\phi}(\mathbf{X}, \mathbf{S})\right)\right), 0\right). \tag{13}$$

So the final $\mathcal{R}$ is

$$\mathcal{R}(S) = \mathbb{E}_{\mathbf{X}\sim P}\left[r\left(\mathbf{X}, S^{\theta}(\mathbf{X})\right)\right]. \tag{14}$$

Inspired by reward-regularized feature selection in [14], we incorporate an entropy regularizer into the selector loss to encourage early exploration and later specialization. The entropy over the selector's soft-assignment $\mathbf{S}$ is

$$\mathcal{H} = -\mathbf{S}\log(\mathbf{S}). \tag{15}$$

Combining this with $r(\cdot)$ from Eq. (14) yields the loss

$$\mathcal{L}_s = \mathbb{E}_{\mathbf{X}\sim P}\left[r\left(\mathbf{X}, S^{\theta}(\mathbf{X})\right) \cdot \mathcal{H}\right] \tag{16}$$

$$= \sum_{\mathbf{x}} \frac{1}{N} r\left(\mathbf{x}, S^{\theta}(\mathbf{x})\right) \cdot (-\mathbf{S}\log(\mathbf{S})). \tag{17}$$

Substituting Eq. (13) gives

$$\mathcal{L}_s = -\frac{1}{N}\sum_{\mathbf{x}} \max\left(m - (b + \sum_{i=1}^{c} y_i \log\left(f_i^{\phi}\left(\mathbf{x}, S^{\theta}(\mathbf{x})\right)\right)\right), 0\right) \cdot \left(S^{\theta}(\mathbf{x}) \log\left(S^{\theta}(\mathbf{x})\right)\right) \tag{18}$$

The selector loss is designed to encourage early exploration and later exploitation, ensuring that the selector network can effectively evaluate and specialize in the most informative expert. Specifically, the optimization process of $\mathcal{L}_s$ ensures that the weight combination $\mathbf{S}$ starts from equal contributions, i.e., [0.5, 0.5]. As $\mathcal{H}$ decreases, $\mathbf{S}$ begins to approach [0, 1] or [1, 0]. This process guarantees exploration of all weight combinations. At this point, $\mathcal{R}$ acts as a performance gate, determining when to cease exploration. Specifically, when $\mathcal{H}$ drives the selector network to explore a particular combination, such as [0.2, 0.8], and $\mathcal{R} = 0$, it indicates that the current weight combination $\mathbf{S}$ meets the requirements, and the gate closes, halting exploration and shifting to exploitation ($\mathcal{L}_s = 0$).

In the context of anomaly detection tasks, when there is a significant discrepancy between the predictions of the two source experts, for instance, if the 3D expert's detection results are inaccurate, to prevent its poor predictions from affecting the final decision, $\mathcal{H}$ will drive the selector network to output a weight combination $\mathbf{S}$ closer to $[0, 1]$. At this time, the performance gate will also close (with $\mathcal{R} = 0$) later, potentially resulting in a weight combination approaching [0, 1], such as [0.1, 0.9], thus assigning a smaller weight to the 3D expert, ensuring that the final prediction is not adversely influenced by errors.

Predictor loss. To ensure that the predictor network can perform the correct nonlinear mapping based on the importance score $\mathbf{S}$ output by the selector, we employ cross-entropy for separate training. This loss function is defined as

$$\mathcal{L}_p = -\frac{1}{N}\sum_{\mathbf{x}}\sum_{i=1}^{c} y_i \log\left(f_i^{\phi}\left(\mathbf{x} \odot S^{\theta}(\mathbf{x})\right)\right). \tag{19}$$

By minimizing this loss, the predictor network learns to effectively combine the weighted score maps into a fused score map $\mathbf{A}$ that accurately reflects the likelihood of anomalies in the input data.

The overall loss function is a combination of the predictor loss $\mathcal{L}_p$ and the selector loss $\mathcal{L}_s$. Specifically, the overall loss $\mathcal{L}_{\text{final}}$ is defined as

$$\mathcal{L}_{\text{final}} = \mathcal{L}_p + \lambda\mathcal{L}_s, \tag{20}$$

where λ is a hyperparameter that balances the trade-off between the predictor loss and the selector loss. This combined loss function ensures that the model not only optimizes the accuracy of the anomaly detection results but also balances the contributions of different experts through the selector network.

Finally, the fusion of the object-level scores $s_1 \in \mathbb{R}$ and $s_2 \in \mathbb{R}$ follows a similar procedure. First, the scores s_1 and s_2 are concatenated

and fed into the same selector network to obtain a weighted score. This weighted score is then passed through the predictor network for nonlinear transformation to produce the final score. Mathematically, this process is expressed as

$$s = f^{\phi}\left(s_1 \oplus s_2,\ S^{\theta}\left(s_1 \oplus s_2\right)\right). \tag{21}$$

The training and inference of our methodology is summarized in Algorithm 1.

Algorithm 1 Model training and inference.

Stage 1: Model training
Input: Normal data $\mathcal{D}$, synthetic data $\mathcal{D}'$, margin m, λ
Output: Dual memory bank $(\mathcal{M}_{2D}, \mathcal{M}_{3D})$, IAF model $\mathcal{F}$
Initialize memory banks $(\mathcal{M}_{2D}, \mathcal{M}_{3D})$
for $\mathbf{P}$ in $\mathcal{D}$ **do**
  $(\mathbf{f}_1, \mathbf{f}_2) \leftarrow$ Extract features from $\mathbf{P}$ using $\mathcal{E}_{3d}$ and $\mathcal{E}_{2d}$.
  Update dual memory bank $(\mathcal{M}_{2D}, \mathcal{M}_{3D})$ with features $(\mathbf{f}_1, \mathbf{f}_2)$.
end for
repeat
  for $\mathbf{P}'$ in $\mathcal{D}'$ **do**
    $(\mathbf{f}'_1, \mathbf{f}'_2) \leftarrow$ Extract features from $\mathbf{P}'$ using $\mathcal{E}_{3d}$ and $\mathcal{E}_{2d}$.
    $(\mathbf{X}_1, \mathbf{X}_2, s_1, s_2) \leftarrow$ Calculate prior anomaly scores using $(\mathcal{M}_{2D}, \mathcal{M}_{3D})$.
    $(\mathbf{X}, s) \leftarrow \mathcal{F}(\mathbf{X}_1, \mathbf{X}_2, s_1, s_2)$.
    Compute final loss Eq. (20) and update $\mathcal{F}$.
  end for
until convergence
Stage 2: Model inference
Input: Testing point cloud $\mathbf{P}_t$
Output: Fused anomaly scores $(\mathbf{A}, s)$
Load the parameters of IAF $\mathcal{F}$ and $(\mathcal{M}_{2D}, \mathcal{M}_{3D})$.
$(\mathbf{f}^t_1, \mathbf{f}^t_2) \leftarrow$ Extract features from $\mathbf{P}_t$ using source experts.
$(\mathbf{X}^t_1, \mathbf{X}^t_2, s^t_1, s^t_2) \leftarrow$ Calculate prior anomaly scores using $(\mathcal{M}_{2D}, \mathcal{M}_{3D})$.
$(\mathbf{A}, s) \leftarrow \mathcal{F}(\mathbf{X}^t_1, \mathbf{X}^t_2, s^t_1, s^t_2)$.

4. Experiments

4.1. Experimental settings

4.1.1. Datasets

Our experiments are conducted on the MVTec 3D-AD [7] industrial dataset and the Eyecandies dataset [29], both of which are commonly used public datasets. MVTec 3D-AD provides real scanned point cloud data for 10 types of objects. For each object category, there are 210 to 361 training samples, all of which are normal data, and 69 to 132 testing samples that include both normal and anomalous data. The anomalous samples encompass 4 to 5 types of anomalies. In summary, the dataset comprises 2656 training samples and 1197 testing samples. Meanwhile, the Eyecandies dataset also covers 10 object categories, with 1000 normal training samples and 50 testing samples for each category (including both normal and anomalous instances).

In addition to real datasets, we also generated synthetic datasets based on both the original MVTec 3D-AD and Eyecandies datasets to train the IAF module. Specifically, we first create an anomaly mask, then utilize a Cut-Paste technique [23] to cut points corresponding to the mask from a source point cloud. The source point cloud could be derived from normal point clouds of other categories or general datasets. These cut points are then pasted onto the target point cloud and subjected to random transformations to produce the synthetic anomaly-enhanced datasets. For each class of objects in both MVTec 3D-AD and Eyecandies datasets, we generated 800 samples to serve as training data.

4.1.2. Implementation details

Data preprocessing. The preprocessing steps for the source expert learning, such as background removal, point cloud patching, and depth map projection, are consistent with the methods described in Shape-Guided [13] and CPMF [33]. Unlike the multi-view projection of CPMF, we only project a frontal view, thus avoiding the choice of projection angle affecting the detection results. These steps are crucial for preparing the data in a format suitable for effective feature extraction and anomaly detection.

Source expert learning. For 3D expert, we adopt the identical architecture of PointNet and SDF in [13]. PointNet outputs patchwise features $\mathbf{f}_1 \in \mathbb{R}^{128}$. Pretraining is performed on the anomaly-free set via SDF reconstruction loss. For 2D expert, we employ a frozen Wide-ResNet-50-2 pretrained on ImageNet. Features from the first two layers are concatenated with FPFH descriptors, yielding an 801 dimensional representation $\mathbf{f}_2$.

IAF training. For the training of our Importance-Aware Fusion (IAF) model, we employ the AdamW optimizer, iterating over 150 epochs with a batch size of 32. The learning rate is set at 0.01, and we utilize a cosine annealing schedule to adjust the learning rate over the training process. In the loss function, we set the margin $m = 0.1$ and the hyperparameter $\lambda = 0.1$. It should be noted that these parameters may require fine-tuning for some categories with significantly varying inputs to balance the contribution of each source models and ensure effective convergence of the model. All experiments are conducted on a single RTX 4090 GPU.

4.1.3. Evaluation metrics

For object-level anomaly detection, we employ the commonly used Area Under the Receiver Operating Characteristic curve (AUROC) metric, denoted as O-AUROC. For point-level anomaly detection, we utilize both the AUROC and the Area Under Per-Region Overlap curve (AUPRO) [7] for evaluation, denoted as P-AUROC and AUPRO, respectively. The AUPRO metric involves calculating the integration of the PRO values over the range of false-positive rates (FPRs). Similar to most previous methods, we set the upper limit of the FPR integration to the default value of 0.3, which is denoted as AUPRO@30%. A smaller FPR integration limit implies a stricter tolerance for false positives.

4.2. Comparison results

MVTec 3D-AD. We compared our approach with other current state-of-the-art 3D anomaly detection methods, including Voxel GAN [7], Voxel AE [7], 3D-ST [10], BTF [11], AST [21], M3DM [24], CPMF [33], and Shape-Guided [13]. To ensure a fair comparison, we also included a variant of CPMF that utilizes only a single view, denoted as CPMF-1view, in our experiments. Table 1 shows the performance of each method for object-level anomaly detection across all categories. Our method ranked second-best, with results very close to the best-performing CPMF (with a difference of just 0.006). The O-AUROC obtained using CPMF-1view are significantly lower than those achieved by our method. However, the anomaly localization performance of CPMF is inferior to that of our approach, as will be illustrated in the subsequent sections. It is worth noting that CPMF relies on 27 viewpoint projections, whose angles and count must be carefully tuned–suboptimal choices can directly degrade performance. In contrast, our 2D expert achieves comparable object-level anomaly detection with just a single projection, avoiding these complications and demonstrating robustness and adaptability in diverse industrial scenarios.

Tables 2 and 3 present the quantitative results for point-level anomaly localization, evaluated by P-AUROC and AUPRO, respectively. It is evident that our method achieves the best results (AUPRO = 0.944), significantly outperforming CPMF (AUPRO = 0.929), which is the top performer in the object-level anomaly detection task. Beyond the superior metrics, the visualization of the score maps in Fig. 5 demonstrates the strengths of our approach. First, the proposed IAENet not only accurately localizes anomalies but also effectively suppresses the anomaly scores in normal regions. This enhances the distinction between anomaly and normal regions.

Another advantage of our IAENet is its capability to effectively detect subtle anomalies on complex surfaces. As depicted in Fig. 5, our

Table 1
O-AUROC scores for anomaly detection across various categories on the MVTec 3D-AD dataset (Best results in bold).

Method	Bagel	Cable gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
Voxel GAN	0.383	0.623	0.474	0.639	0.564	0.409	0.617	0.427	0.663	0.577	0.537
Voxel AE	0.693	0.425	0.515	0.790	0.494	0.558	0.537	0.484	0.639	0.583	0.571
BTF (SIFT)	0.711	0.656	0.892	0.754	0.828	0.686	0.622	0.754	0.767	0.598	0.727
BTF (FPFH)	0.825	0.551	0.952	0.797	0.883	0.582	0.758	0.889	0.929	0.653	0.782
AST	0.881	0.576	0.956	0.957	0.679	0.797	0.990	0.915	0.956	0.611	0.832
M3DM	0.941	0.651	0.965	0.969	0.905	0.760	0.880	0.974	0.926	0.765	0.874
CPMF	0.981	0.888	**0.992**	0.989	0.962	**0.794**	0.990	0.963	**0.979**	0.966	**0.950**
CPMF-1view	0.957	**0.969**	0.984	0.831	0.887	0.707	0.931	**0.975**	0.876	0.927	0.904
Shape-Guided	**0.983**	0.710	0.974	**0.993**	**0.971**	0.722	**0.992**	0.964	0.966	0.931	0.921
Ours	0.969	0.954	0.990	0.945	0.955	0.756	0.983	0.966	0.954	**0.968**	0.944

Table 2
P-AUROC scores for anomaly localization across various categories on the MVTec 3D-AD dataset (Best results in bold).

Method	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
BTF (SIFT)	0.974	0.862	0.993	0.952	**0.980**	0.862	0.955	0.996	0.993	0.971	0.954
BTF (FPFH)	**0.995**	0.965	**0.999**	0.947	0.966	0.928	0.996	**0.999**	**0.996**	0.991	0.978
M3DM	0.981	0.947	0.996	0.934	0.960	**0.944**	0.988	0.994	0.994	0.983	0.972
CPMF	0.986	0.986	0.997	0.926	0.969	0.936	0.997	0.998	**0.996**	0.996	0.979
CPMF-1vew	0.976	0.983	0.996	0.910	0.961	0.934	0.997	0.998	0.991	0.991	0.974
Shape-Guided	0.991	0.962	0.998	0.947	0.959	0.930	0.996	**0.999**	0.995	0.996	0.977
Ours	0.994	**0.988**	0.998	**0.964**	0.944	**0.944**	**0.998**	**0.999**	0.994	**0.997**	**0.982**

Table 3
AUPRO@30% scores for anomaly localization across various categories on the MVTec 3D-AD dataset (Best results in bold).

Method	Bagel	Cable gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
Voxel GAN	0.440	0.453	0.875	0.755	0.782	0.378	0.392	0.639	0.775	0.389	0.583
Voxel AE	0.260	0.341	0.581	0.351	0.502	0.234	0.351	0.658	0.015	0.185	0.348
3D-ST	0.950	0.483	**0.986**	0.921	0.905	0.632	0.945	**0.988**	**0.976**	0.542	0.833
BTF (SIFT)	0.942	0.842	0.974	0.896	**0.910**	0.723	0.944	0.981	0.953	0.929	0.909
BTF (FPFH)	0.973	0.879	0.982	0.906	0.892	0.735	0.977	0.982	0.956	0.961	0.924
M3DM	0.943	0.818	0.977	0.882	0.881	0.743	0.958	0.974	0.950	0.929	0.906
CPMF	0.958	0.946	0.979	0.868	0.897	0.746	0.980	0.981	0.961	0.977	0.929
CPMF-1view	0.940	0.937	0.977	0.826	0.879	0.741	0.978	0.982	0.939	0.962	0.916
Shape-Guided	0.972	0.867	0.981	0.922	0.897	0.767	0.978	0.983	0.955	0.972	0.929
Ours	**0.975**	**0.951**	0.982	**0.946**	0.877	**0.814**	**0.982**	0.983	0.956	**0.978**	**0.944**

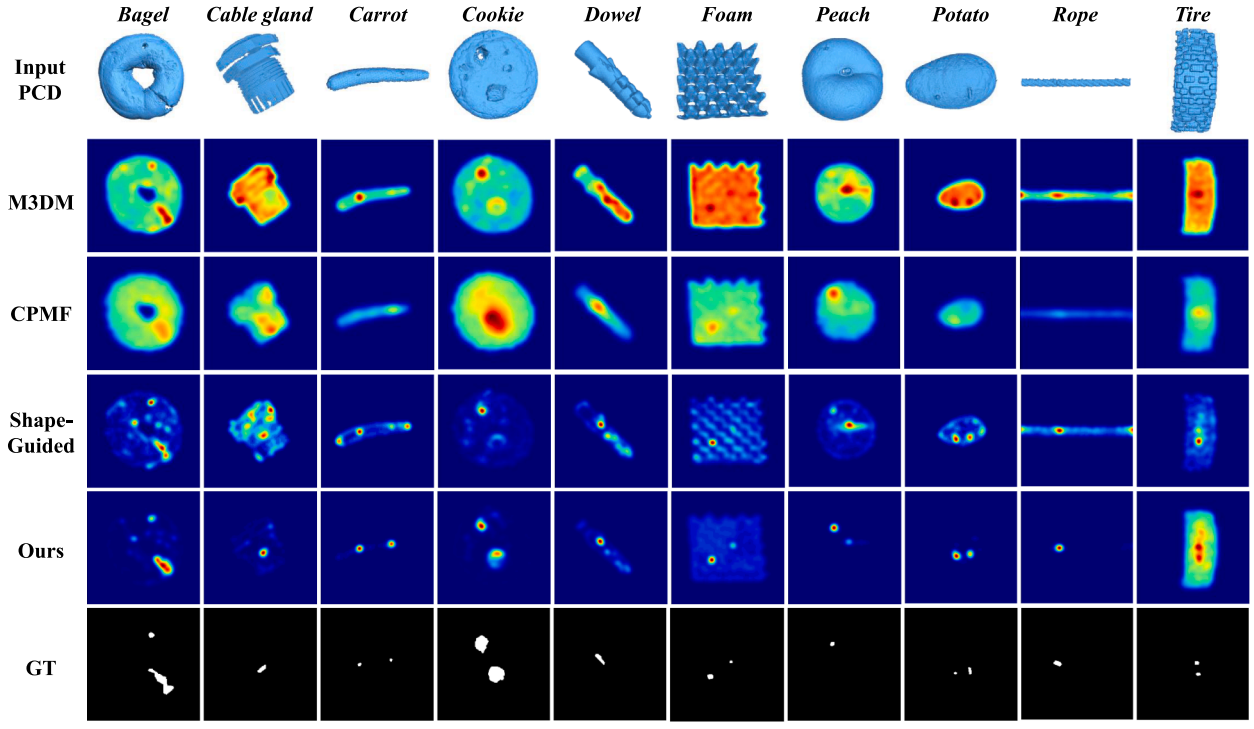

Fig. 5. Qualitative comparison of anomaly score maps on the MVTec 3D-AD dataset. This is the visualization of the final output of the anomaly score map for each methodology, where blue represents values close to 0 and red represents larger anomaly score values.

approach effectively identifies the scratch on the thread surface of a cable gland, an anomaly that other methods struggle to identify. The score map produced by our method distinctly marks the location of the anomaly, with anomaly scores for normal regions being close to zero. This demonstrates that our IAENet not only accurately locates subtle anomalies with high confidence but also accurately discerns complex normal surfaces.

Fig. 6 displays the object-level anomaly detection and point-level anomaly localization results by the 2D expert, 3D expert, and our IAENet on the MVTec 3D-AD dataset. It is evident that across both object-level anomaly detection and point-level anomaly localization, IAENet (with mean O-AUROC of 0.944 and mean AUPRO of 0.944) outperforms individual expert models across all categories: the 2D expert achieves mean O-AUROC of 0.904 and mean AUPRO of 0.914, while the 3D expert attains mean O-AUROC of 0.921 and mean AUPRO of 0.929. Furthermore, it is clear that our IAF module effectively manages scenarios with significant discrepancies in the results of different sources of experts. For instance, in the object-level anomaly detection for the *Cable gland* category, where the 2D expert scored 0.969 and the 3D expert scored 0.71, the IAF module still achieves a satisfactory result of 0.954, unaffected by the inferior performance. This confirms the efficacy of the proposed IAF module, demonstrating its ability to determine which source expert's results contribute more favorably to the final anomaly detection.

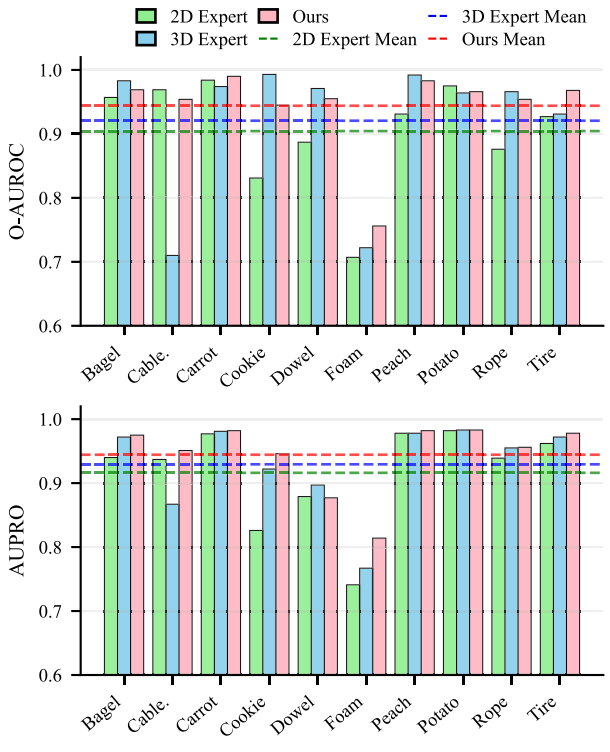

Fig. 6. Object-level (O-AUROC) and point-level (AUPRO) anomaly detection results on the MVTec 3D-AD dataset by the 2D expert, 3D expert, and our IAENet.

Eyecandies. We compared our approach with other current state-of-the-art 3D anomaly detection methods on the Eyecandies dataset, including BTF [11], AST [21], M3DM [24], CPMF [33], Shape-Guided [13], EasyNet [45] and DMVAD [46]. As shown in Tables 4 and 5, the results demonstrate that our method achieves the best performance in both object-level and point-level anomaly detection. The visualization results in Fig. 7 also exhibit similar performance to those on the MVTec 3D-AD dataset, thereby further validating the effectiveness of our approach.

In addition, the comparison of the results between the 2D expert, 3D expert, and IAENet is illustrated in Fig. 8. The results indicate that IAENet consistently outperforms either individual expert model. However, unlike the results on the MVTec 3D-AD dataset, the performance of IAENet on the AUPRO metric is not significantly higher than that of the 2D expert. This is attributed to the fact that the 2D expert achieves substantially better results than the 3D expert on the Eyecandies dataset, thereby providing limited complementary information to our IAF module. Consequently, the final results of IAENet are only marginally better than those of the 2D expert. It is worth noting that the Eyecandies dataset provides accurate normal maps for each object. We utilized these normal maps to replace our normal-based rendering similar to [46]. With the more precise depth maps derived from the accurate normal maps, the 2D expert exhibits superior performance on the Eyecandies dataset.

4.3. Analysis

4.3.1. Difference between anomaly and normal points

As demonstrated by the visualization results, one of the strengths of our IAENet is its ability to effectively suppress the anomaly scores of normal points, thereby magnifying the difference between anomalies and normal regions. In this section, we will quantify and analyze this advantage by examining the results of AUPRO at different integration limits and the distribution of anomaly scores. Note that all analysis experiments are conducted on the MVTec 3D-AD dataset.

Comparison on different integration limits. Our previous AUPRO metrics are conducted with an integration upper limit of 0.3, i.e., AUPRO@30%, as mentioned in Section 4.1.3, where a smaller integration limit indicates a stricter tolerance for false positives. Therefore, we compared our IAENet with benchmarks across seven decreasing integration limits $\{0.3, 0.2, 0.1, 0.07, 0.05, 0.03, 0.01\}$ for the AUPRO results, as shown in Fig. 9. It can be observed that across the various integration limits, our method consistently achieves the state-of-the-art performance.

Anomaly score distributions. Fig. 10 illustrates the distribution of anomaly scores for CPMF and our method across three categories. It is evident that CPMF exhibits higher anomaly scores in normal regions, and the density of anomaly scores near zero is significantly lower than that of our method. For instance, in the *Bagel* category, the CPMF has a density of 6, while our IAENet has a density of 40. In contrast, our approach effectively suppresses the anomaly scores of normal points, as evidenced by the anomaly scores for the majority of normal points being close to zero. This property enables our method to reduce the false positive rate, allowing it to outperform comparative methods even at lower integration limits (as shown in Fig. 9), and thus has greater practical value in industrial applications.

4.3.2. Comparison of fusion strategies

To validate the effectiveness of the proposed IAF module, we compare it with existing score map fusion strategies, including pixel-wise addition [11], maximum selection [13], and data-driven one-class Support Vector Machine (OCSVM) [24]. Table 6 presents the quantitative results of different fusion strategies, demonstrating that our IAF module consistently achieves the best performance across various metrics. This comparison underscores the superiority of our approach in effectively integrating the strengths of multiple models to enhance anomaly detection accuracy.

The visualization results for different fusion strategies are shown in Fig. 11. As we can see, the 3D expert tends to identify significant anomalies with high confidence, as indicated by the large difference between its normal and anomaly scores. Conversely, the 2D expert has a tendency to detect all potential anomalies, which is reflected in relatively high normal scores. Based on the outputs from the source experts, our IAF module not only integrates the collective knowledge of the source experts but also preserves their unique strengths. This means it can identify all potential anomalies with high confidence. In contrast, other fusion strategies fail to achieve this balance. Addition and Max strategies

Table 4
O-AUROC scores for anomaly detection across various categories on the Eyecandies dataset (Best results in bold).

Method	Candy Cane	Chocolate Cookie	Chocolate Praline	Confetto	Gummy Bear	Hazelnut Truffle	Licorice Sandwich	Lollipop	Marsh-mallow	Peppermint Candy	Mean
EasyNet	0.629	0.716	0.768	0.731	0.660	**0.710**	0.712	0.711	0.688	0.731	0.706
M3DM	0.482	0.589	0.805	0.845	0.780	0.538	0.766	0.827	0.800	0.822	0.725
BTF	0.670	0.728	0.806	0.806	0.721	0.514	0.794	0.757	0.758	0.757	0.731
CPMF	0.773	0.795	0.750	0.846	0.740	0.584	0.746	0.820	0.710	0.813	0.758
Shape.	0.408	0.656	0.746	0.837	0.752	0.643	0.776	0.825	0.805	0.755	0.720
DMVAD	0.790	0.885	**0.933**	**0.915**	0.837	0.517	0.888	**0.960**	**0.941**	**0.949**	0.862
Ours	**0.891**	**0.906**	0.854	0.891	**0.860**	0.643	**0.899**	0.935	0.882	0.910	**0.867**

Table 5
AUPRO@30% scores for anomaly localization across various categories on the Eyecandies dataset (Best results in bold).

Method	Candy Cane	Chocolate Cookie	Chocolate Praline	Confetto	Gummy Bear	Hazelnut Truffle	Licorice Sandwich	Lollipop	Marsh-mallow	Peppermint Candy	Mean
BTF	0.944	0.725	0.687	0.601	0.651	0.471	0.636	0.885	0.598	0.594	0.679
M3DM	0.911	0.645	0.581	0.748	0.748	0.484	0.608	0.904	0.646	0.750	0.702
CPMF	0.941	0.756	0.584	0.795	0.728	0.561	0.679	0.908	0.696	0.784	0.743
Shape.	0.923	0.671	0.583	0.688	0.653	0.574	0.672	0.927	0.641	0.682	0.701
DMVAD	0.842	**0.841**	**0.870**	0.868	0.814	0.591	0.838	0.865	0.776	0.774	0.808
Ours	**0.957**	0.804	0.757	**0.932**	**0.857**	**0.621**	**0.854**	**0.952**	**0.818**	**0.915**	**0.847**

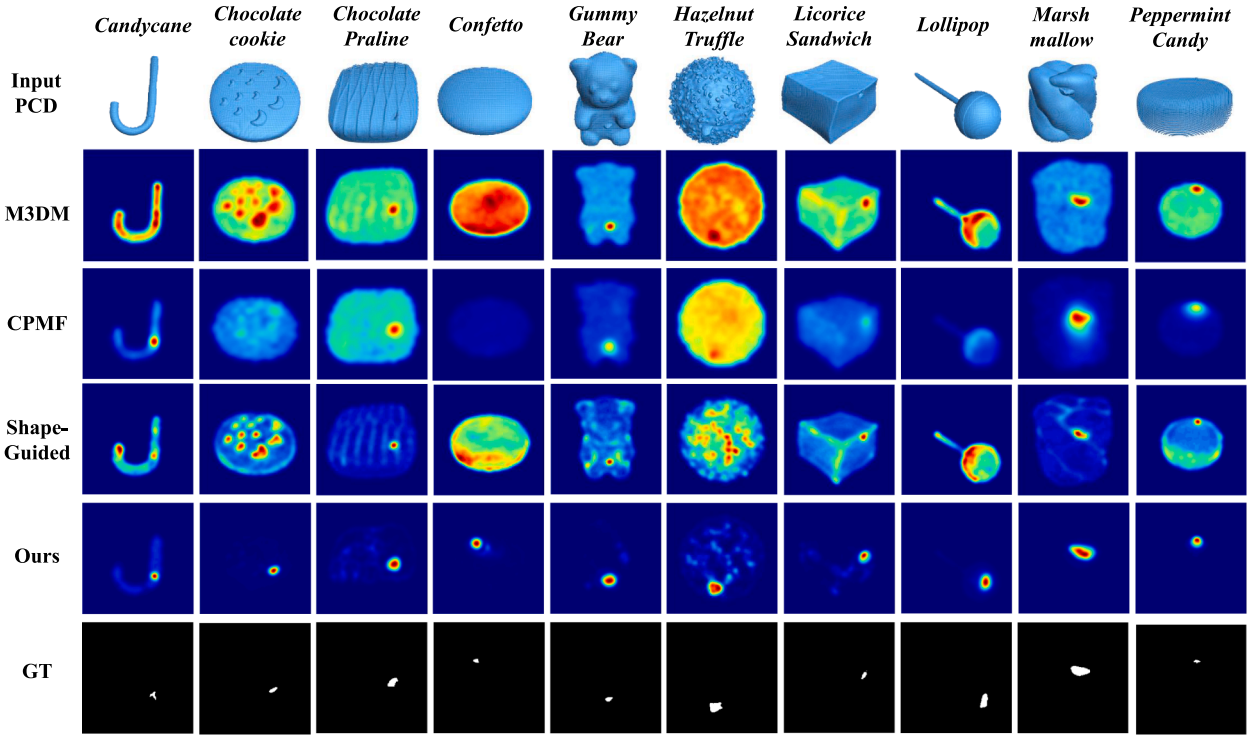

Fig. 7. Qualitative comparison of anomaly score maps on the Eyecandies dataset. This is the visualization of the final output of the anomaly score map for each methodology, where blue represents values close to 0 and red represents larger anomaly score values.

Table 6
Mean values of different fusion strategies across all categories of the MVTec 3D-AD dataset.

Stategies	O-AUROC	P-AUROC	AURPO@30%	AUPRO@1%
Max	0.921	0.974	0.916	0.367
Addition	0.934	0.975	0.920	0.383
OCSVM	0.921	0.977	0.929	0.399
Ours	**0.944**	**0.982**	**0.944**	**0.424**

are biased towards the 2D expert due to the higher values in its anomaly maps, while OCSVM favors the 3D expert, which performs better in the anomaly detection task. However, OCSVM cannot effectively utilize information from the 2D expert, potentially missing some anomalies, as illustrated in the *Cookie* example in the Fig. 11.

4.3.3. Ablation study

The IAF module is composed of two main components: the selector network S^{θ} and the predictor network f^{ϕ}. The selector network is primarily responsible for evaluating the contributions of different experts to the final anomaly detection task and providing corresponding importance scores. The predictor network, on the other hand, performs a nonlinear mapping on the weighted score maps. The selector network is trained using a specially designed loss function, $\mathcal{L}_s$. Therefore, in this section, we will examine the impact of the three critical components on the final anomaly detection performance: S^{θ}, f^{ϕ}, and $\mathcal{L}_s$.

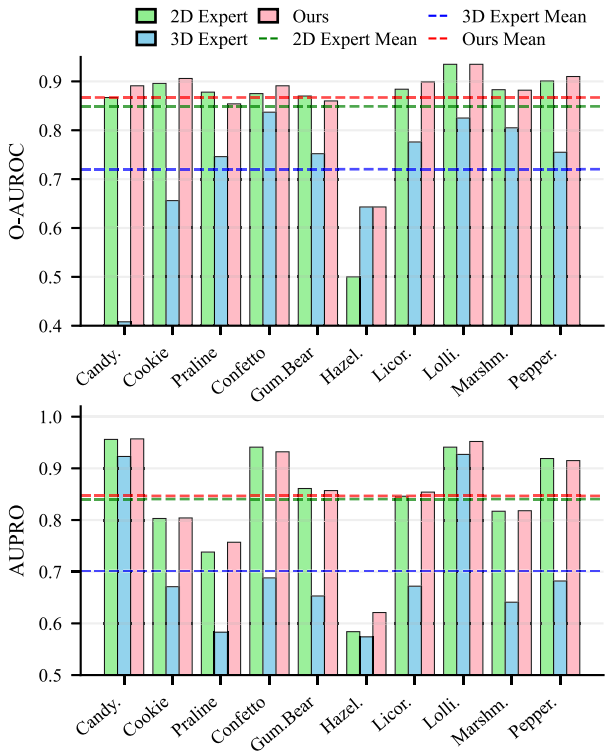

Fig. 8. Object-level and point-level anomaly detection results on the Eyecandies dataset by the 2D expert, 3D expert, and our IAENet.

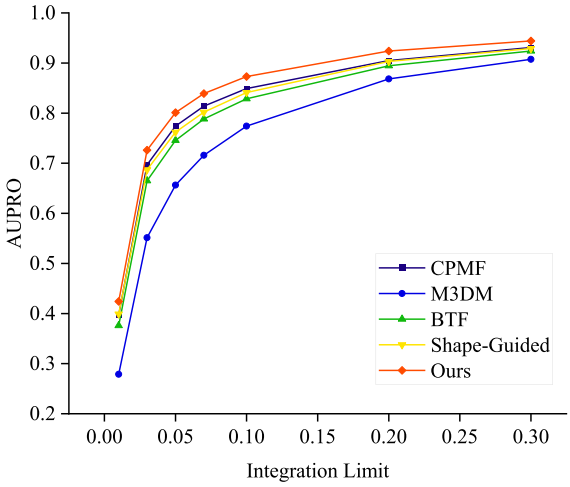

Fig. 9. Anomaly localization performance (AURPO) of our method with comparison methods at different integration limits.

It can be observed in Table 7 that when only the predictor network f^{ϕ} is present, indicating a simple nonlinear mapping, the results are the poorest because it treats the outputs from different experts equally. Upon adding the selector network S^{θ}, the model attempts to assess the contributions of various experts, but is solely optimized by the end-to-end anomaly detection loss, which fails to yield reliable importance scores. However, with the inclusion of $\mathcal{L}_s$, the selector network is explicitly guided in its optimization, enabling it to more accurately evaluate the contributions of different experts, thereby achieving the best anomaly detection results. It is noteworthy that if only $\mathcal{L}_s$ is used without $\mathcal{H}$, the model also fails to achieve satisfactory results, indicating that $\mathcal{H}$ plays a significant role in obtaining the optimal combination of weights for the final decision.

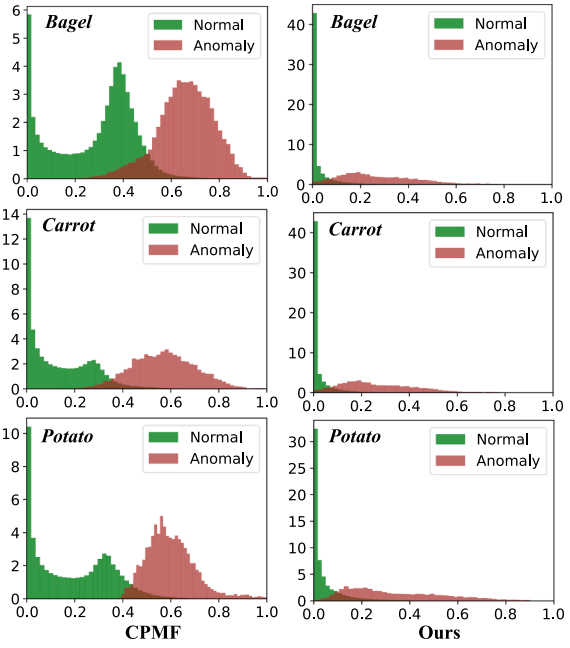

Fig. 10. Point-level anomaly score distributions across three categories. The x-axis represents anomaly scores, and the y-axis represents probability density. It can be observed that our approach effectively suppresses the anomaly scores in normal regions compared to CPMF.

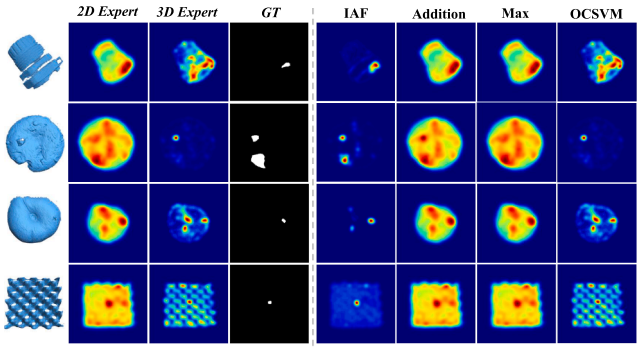

Fig. 11. Qualitative comparison of different fusion strategies in representative categories, including *Cable gland, Cookie, Peach and Foam*. The left side shows the input point clouds, the fusion inputs and ground truth, and the right side shows the outputs of different fusion strategies.

Table 7
Ablation study of our IAENet.

Components	Stepwise Combinations			
Predictor Network f^{ϕ}	✓	✓	✓	✓
Selector Network S^{θ}		✓	✓	✓
$\mathcal{L}_s$ w/o $\mathcal{H}$			✓	✓
$\mathcal{L}_s$ w/ $\mathcal{H}$				✓
O-AUROC	0.938	0.942	0.934	**0.944**
P-AUROC	0.966	0.970	0.971	**0.982**
AURPO@30%	0.921	0.924	0.925	**0.944**

4.3.4. Investigation of IAF module

In our IAF module, the selector network is designed to evaluate and select which of the 2D expert or 3D expert contributes more to the final anomaly scores. Fig. 12 illustrates the behavior of the selector and predictor networks. The 2D score map $\mathbf{X}_2$ and 3D score map $\mathbf{X}_1$ serve as the inputs to the selector network, while the selection map visualizes the selection behavior of the selector network, defined as $(\mathbf{X}_2 \odot \mathbf{S}_2 - \mathbf{X}_1 \odot \mathbf{S}_1)$. Specifically, regions closer to red indicate that the selector network

favors the 2D expert at that location, whereas regions closer to blue indicate a preference for the 3D expert. We do not use $(\mathbf{S}_2 - \mathbf{S}_1)$ alone because $\mathbf{S}_1$ and $\mathbf{S}_2$ depend on their corresponding score maps; shown alone they do not reflect the selector's effect on the final scores. The selector output is defined as $(\mathbf{X}_1 \odot \mathbf{S}_1 + \mathbf{X}_2 \odot \mathbf{S}_2)$, representing the linear combination of the score maps after weighting. The predictor output is the result after the non-linear transformation by the predictor network.

As shown in Fig. 12, the inputs to the selector network exhibit a certain degree of complementarity. The selection map demonstrates that our method effectively focuses on and selects the more contributive expert model results at different locations. The selector output reveals that our selector network has successfully integrated the complementary information from different experts. Finally, through the non-linear mapping of the predictor network, we obtain more accurate and discriminative results.

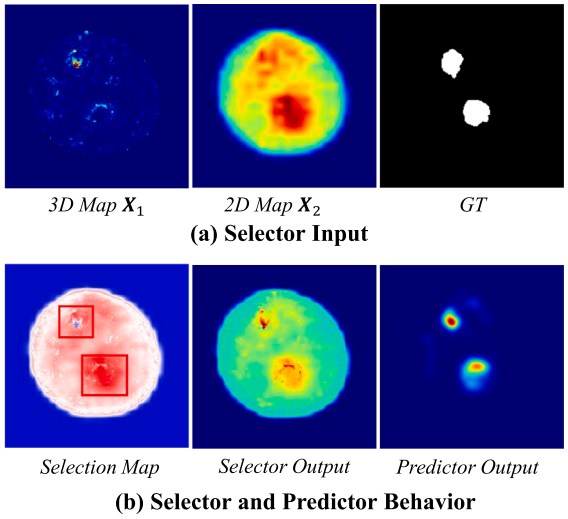

Fig. 12. Visualization of the behavior of the selector and predictor networks in the IAF module. **(a) Selector Input**. The 2D Score Map and 3D Score Map are the inputs to the selector network. **(b) Selector and Predictor Behavior**. The Selection Map visualizes the selector network's choices, with red indicating preference for the 2D expert and blue indicating preference for the 3D expert. The Selector Output shows the weighted linear combination of the score maps, while the Predictor Output represents the final results after non-linear transformation by the predictor network.

4.3.5. Sensitivity analysis

We conducted a sensitivity analysis on two critical parameters, λ and m, as illustrated in Fig. 13. The margin m is notably more sensitive for object-level detection while remaining stable for point-level. O-AUROC stays constant for m in [0.1, 0.25]; outside this interval O-AUROC degrades. This behavior is expected: a too-small margin does not sufficiently encourage the fused score to exceed individual experts' outputs, causing training to stall and preventing learning of finer distinctions; conversely, an excessively large margin enforces an impractically large separation, which hinders optimization and can lead to overfitting.

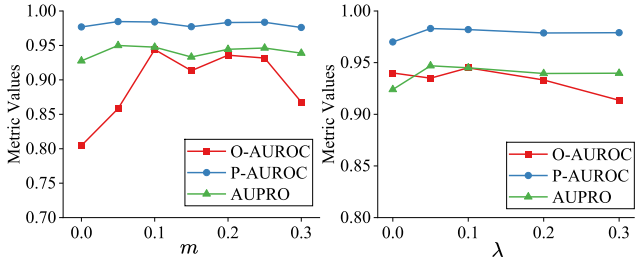

Fig. 13. Sensitivity analysis of parameters λ and m.

Similarly, the parameter λ initially leads to an increase in point-level detection performance after surpassing 0, followed by demonstrating stability across a broader range of values. In contrast, for object-level detection, the O-AUROC shows some variation, reaching its maximum at $\lambda = 0.1$, and then begins to gradually decrease as λ exceeds 0.2. Consequently, we have selected $\lambda = 0.1$ as the default hyperparameter for our method.

4.3.6. Investigation of different synthetic samples

Overfitting to synthetic artifacts is an inherent concern when training on artificially generated anomalies. To evaluate the risk of overfitting to Cut-Paste artifacts, we conduct experiments by varying the number of training samples, as shown in Table 8. The results indicate that performance only slightly decreases with fewer samples, suggesting that our model does not overfit to these artifacts. Instead, it learns to differentiate between normal and anomalous distributions rather than memorizing the specific patterns of anomalies, thus minimizing the impact.

Table 8
Anomaly detection results under different numbers of synthetic training samples.

Training samples	O-AUROC	P-AUROC	AUPRO
800	0.944	0.982	0.944
600	0.929	0.979	0.939
400	0.941	0.976	0.932
200	0.935	0.976	0.936

4.3.7. Efficiency comparison

Table 9 presents a detailed analysis of our IAENet's GPU memory usage and computational efficiency. Inference time and frame per second (FPS) are used to measure performance, with higher values indicating better efficiency. Memory usage is assessed by the amount of GPU memory occupied during inference, with lower values being more desirable.

Table 9
Comparison of GPU memory usage, frame per second (FPS), and inference time per sample on MVTec 3D-AD. The upper panel reports each step for our method (2D expert, 3D expert, and the IAF fusion module); the lower panel compares the end-to-end IAENet with baseline methods.

Method	Inference Time (s)	FPS	Memory Usage (MB)
2D expert	0.140	7.36	1045
3D expert	0.766	1.30	196.3
IAF module	0.002	305	27.29
IAENet	0.925	1.08	1090
Shape-Guided	0.766	1.30	196.3
CPMF	0.240	3.97	5525
M3DM	1.240	0.81	1997

Our model's inference time is measured without considering offline preprocessing and memory bank construction. IAENet outperforms M3DM in total runtime and is slightly slower than Shape-Guided. The IAF fusion module processes each sample in 0.002 s, contributing minimally to the overall time consumption. The 2D expert, using a single-view image, is nearly twice as fast as CPMF, which uses 27 views.

In terms of memory usage, IAENet requires slightly more than Shape-Guided. The 2D expert, with its Wide-ResNet-50-2 model (4.13M parameters), is the primary consumer of memory. The 3D expert (3.17M parameters) and the IAF module (0.018M parameters) add minimal overhead. However, despite using a larger model than CPMF's ResNet18, our method is more memory-efficient due to the single-view strategy. The 3D expert complements this by providing 3D information, ensuring high

performance without excessive memory demands. The low memory requirement will be very beneficial for edge deployment, which will be more practical for the real industrial applications.

4.3.8. Multimodal anomaly detection

Furthermore, given that our IAENet is capable of accepting RGB inputs through the 2D expert, we have also evaluated its potential on multimodal anomaly detection tasks using the MVTec 3D-AD dataset. Specifically, to handle multimodal inputs, we no longer need to project the point clouds. Instead, we replace the 2D expert's input with RGB images, while keeping the 3D expert unchanged. In this experiment, we have included the state-of-the-art multimodal anomaly detection method LSFA [47] for comparison. As shown in Table 10, our IAENet achieves the best results in object-level anomaly detection. In terms of AUPRO@30%, our method ranks second, comparable to LSFA and just behind Shape-Guided. Notably, on the AUPRO@1% metric IAENet matches Shape-Guided, reflecting effective suppression of spurious scores in normal regions and enabling more confident anomaly localization with a lower false-positive rate in multimodal detection. In summary, while our method is primarily designed for 3D anomaly detection, it also shows promise in multimodal anomaly detection tasks, and it holds significant potential for practical applications where minimizing false positives is crucial.

Table 10
Multimodal anomaly detection results on the MVTec 3D-AD dataset.

Method	O-AUROC	P-AUROC	AURPO@30%	AURPO@1%
BTF	0.865	0.992	0.959	0.383
AST	0.937	0.976	0.944	0.398
M3DM	0.945	0.992	0.964	0.393
Shape.	0.947	**0.996**	**0.976**	**0.456**
LSFA	0.971	0.993	0.968	-
Ours	**0.974**	0.992	0.967	**0.456**

5. Conclusion

In this work we present IAENet, a novel ensemble framework that seamlessly integrates a 2D pretrained foundation model with a dedicated 3D expert to push the frontiers of 3D point cloud-based anomaly detection. Recognizing that naive fusion of heterogeneous predictions can be undermined by large inter-expert discrepancies, we introduce the Importance-Aware Fusion (IAF) module together with specifically designed loss functions. IAF dynamically reweights the contribution of each model, ensuring that the final decision leverages their complementary strengths while preserving their individual merits. Extensive experiments on MVTec 3D-AD show that IAENet achieves a new state-of-the-art for point-level localization and ranks second at object level. On the Eyecandies dataset, IAENet attains state-of-the-art performance at both point and object levels. Ablation studies further validate the necessity of every component within IAF. Notably, our method effectively suppresses anomaly scores on normal regions, yielding markedly lower false positive rates, an attribute of critical value for real-world industrial deployment.

CRediT authorship contribution statement

Xuanming Cao: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization; **Chengyu Tao:** Writing – review & editing, Formal analysis, Data curation; **Yifeng Cheng:** Writing – review & editing, Investigation; **Juan Du:** Writing – review & editing, Supervision, Project administration.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China 72371219 and 72001139, and the Guangzhou Municipal Science and Technology Project Guangzhou Municipal Science and Technology Project 2025A04J5288.

References

[1] X. Cao, C. Tao, J. Du, 3D-ADCS: untrained 3D anomaly detection for complex manufacturing surfaces, J. Comput. Inf. Sci. Eng. 25 (11) (2025) 111002. https://doi.org/10.1115/1.4068472
[2] J. Du, C. Tao, X. Cao, F. Tsung, 3D vision-based anomaly detection in manufacturing: a survey, Front. Eng. Manage. 12 (2025) 343–360. https://doi.org/10.1007/s42524-025-4189-9
[3] Y. Lin, Y. Chang, X. Tong, J. Yu, A. Liotta, G. Huang, W. Song, D. Zeng, Z. Wu, Y. Wang, et al., A survey on RGB, 3D, and multimodal approaches for unsupervised industrial image anomaly detection, Inf. Fusion 121 (2025) 103139. https://doi.org/10.1016/j.inffus.2025.103139
[4] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, P. Gehler, Towards total recall in industrial anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14318–14328.
[5] T. Defard, A. Setkov, A. Loesch, R. Audigier, Padim: a patch distribution modeling framework for anomaly detection and localization, in: International Conference on Pattern Recognition, Springer, 2021, pp. 475–489.
[6] M. Ntoulmperis, S. Discepolo, P. Castellini, P. Catti, N. Nikolakis, W. van de Kamp, K. Alexopoulos, Inline-acquired product point clouds for non-Destructive testing: a case study of a steel part manufacturer, Machines 13 (2) (2025) 88.
[7] P. Bergmann, X. Jin, D. Sattlegger, C. 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, 5, 2022, pp. 202–213.
[8] W. Dong, K. Sheng, B. Huang, K. Xiong, K. Liu, X. Cheng, Stretchable self-powered TENG sensor array for human robot interaction based on conductive ionic gels and LSTM neural network, IEEE Sens. J. 24 (22) (2024) 37962–37969. https://doi.org/10.1109/JSEN.2024.3464633
[9] Y. Regaya, F. Fadli, A. Amira, 3D point cloud enhancement using unsupervised anomaly detection, in: 2019 International Symposium on Systems Engineering (ISSE), IEEE, 2019, pp. 1–6.
[10] P. Bergmann, D. Sattlegger, Anomaly detection in 3D point clouds using deep geometric descriptors, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2613–2623.
[11] E. Horwitz, Y. 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, 2023, pp. 2968–2977.
[12] R.B. Rusu, N. Blodow, M. Beetz, Fast point feature histograms (FPFH) for 3D registration, in: 2009 IEEE International Conference on Robotics and Automation, IEEE, 2009, pp. 3212–3217.
[13] Y.-M. Chu, C. Liu, T.-I. Hsieh, H.-T. Chen, T.-L. Liu, Shape-Guided dual-memory learning for 3D anomaly detection, in: Proceedings of the 40th International Conference on Machine Learning, 2023, pp. 6185–6194.
[14] J. Yoon, J. Jordon, M. Van der Schaar, INVASE: instance-wise variable selection using neural networks, in: International Conference on Learning Representations, 2018.
[15] X. Liu, C. Wu, H. Zhang, L. Wang, Memory association guided unsupervised anomaly detection with adaptive 3D attention, Inf. Fusion 124 (2025) 103379. https://doi.org/10.1016/j.inffus.2025.103379
[16] Z. Li, Y. Ge, L. Meng, A multi-scale information fusion framework with interaction-aware global attention for industrial vision anomaly detection and localization, Inf. Fusion 124 (2025) 103356. https://doi.org/10.1016/j.inffus.2025.103356
[17] P. Bergmann, M. Fauser, D. Sattlegger, C. Steger, MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9592–9600.
[18] D. Gudovskiy, S. Ishizaka, K. Kozuka, CFLOW-AD: real-time unsupervised anomaly detection with localization via conditional normalizing flows, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 98–107.
[19] J. Bae, J.-H. Lee, S. Kim, PNI: industrial anomaly detection using position and neighborhood information, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 6373–6383.
[20] P. Bergmann, M. Fauser, D. Sattlegger, C. Steger, Uninformed students: student-teacher anomaly detection with discriminative latent embeddings, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4183–4192.
[21] M. Rudolph, T. Wehrbein, B. Rosenhahn, B. Wandt, Asymmetric student-teacher networks for industrial anomaly detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2592–2602.

[22] X. Zhang, S. Li, X. Li, P. Huang, J. Shan, T. Chen, DeSTSeg: segmentation guided denoising student-teacher for anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3914–3923.
[23] C. Tao, X. Cao, J. Du, G^{2}SF-MIAD: geometry-guided score fusion for multimodal industrial anomaly detection, arXiv preprint arXiv:2503.10091, (2025).
[24] Y. Wang, J. Peng, J. Zhang, R. Yi, Y. Wang, C. Wang, Multimodal industrial anomaly detection via hybrid fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8032–8041.
[25] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, ICLR (2021).
[27] Y. Cao, Q. Wan, W. Shen, L. Gao, Informative knowledge distillation for image anomaly segmentation, Knowl. Based Syst. 248 (2022) 108846.
[28] M.B. Ammar, A. Mendoza, N. Belkhir, A. Manzanera, G. Franchi, Foundation models and transformers for anomaly detection: a survey, Inf. Fusion 126 (2026) 103517. https://doi.org/10.1016/j.inffus.2025.103517
[29] L. Bonfiglioli, M. Toschi, D. Silvestri, N. Fioraio, D. De Gregorio, The eyecandies dataset for unsupervised multimodal anomaly detection and localization, in: Proceedings of the Asian Conference on Computer Vision, 2022, pp. 3586–3602.
[30] J. Liu, G. Xie, R. Chen, X. Li, J. Wang, Y. Liu, C. Wang, F. Zheng, Real3D-AD: a dataset of point cloud anomaly detection, Adv. Neural Inf. Process. Syst. 36 (2023) 30402–30415.
[31] C.R. Qi, H. Su, K. Mo, L.J. Guibas, PointNet: deep learning on point sets for 3D classification and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 652–660. Honolulu, HI, USA
[32] Y. Pang, W. Wang, F.E.H. Tay, W. Liu, Y. Tian, L. Yuan, Masked autoencoders for point cloud self-supervised learning, in: European Conference on Computer Vision, Springer, 2022, pp. 604–621.
[33] Y. Cao, X. Xu, W. Shen, Complementary pseudo multimodal feature for point cloud anomaly detection, Pattern Recognit. 156 (2024) 110761.
[34] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, H. Li, PointCLIP: point cloud understanding by clip, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8552–8562.
[35] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, A. Markham, RandLa-Net: efficient semantic segmentation of large-scale point clouds, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11108–11117.
[36] W. Li, X. Xu, Y. Gu, B. Zheng, S. Gao, Y. Wu, 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, 2024, pp. 22207–22216.
[37] V. Zavrtanik, M. Kristan, D. Skočaj, Cheating depth: enhancing 3D surface anomaly detection via depth simulation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 2164–2172.
[38] M. Masuda, R. Hachiuma, R. Fujii, H. Saito, Y. Sekikawa, Toward unsupervised 3D point cloud anomaly detection using variational autoencoder, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 2021, pp. 3118–3122.
[39] Q. Zhou, J. Yan, S. He, W. Meng, J. Chen, PointAD: comprehending 3d anomalies from points and pixels for zero-shot 3d anomaly detection, Adv. Neural Inf. Process. Syst. 37 (2024) 84866–84896.
[40] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[41] C. Tao, J. Du, PointSGRADE: sparse learning with graph representation for anomaly detection by using unstructured 3D point cloud data, IISE Trans. 57 (2) (2025) 131–144.
[42] A.X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al., Shapenet: An information-rich 3D model repository, arXiv preprint arXiv:1512.03012 (2015).
[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[44] F. Wan, X. Huang, D. Cai, X. Quan, W. Bi, S. Shi, Knowledge fusion of large language models, in: The Twelfth International Conference on Learning Representations, 2024.
[45] R. Chen, G. Xie, J. Liu, J. Wang, Z. Luo, J. Wang, F. Zheng, Easynet: an easy network for 3D industrial anomaly detection, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 7038–7046.
[46] C. Liu, Y.-M. Chu, T.-I. Hsieh, H.-T. Chen, T.-L. Liu, Learning diffusion models for multi-view anomaly detection, in: European Conference on Computer Vision, Springer, 2024, pp. 328–345.
[47] Y. Tu, B. Zhang, L. Liu, Y. Li, J. Zhang, Y. Wang, C. Wang, C. Zhao, Self-supervised feature adaptation for 3D industrial anomaly detection, in: European Conference on Computer Vision, Springer, 2024, pp. 75–91.