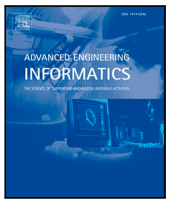
Advanced Engineering Informatics 65 (2025) 103240

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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



Full length article

CPIR: Multimodal Industrial Anomaly Detection via Latent Bridged Cross-modal Prediction and Intra-modal Reconstruction

Wen Shangguan [a,b,1], Hongqiang Wu [c,1], Yanchang Niu [a], Haonan Yin [a], Jiawei Yu [a,b], Bokui Chen [b], Biqing Huang [a,*]

[a] Department of Automation, Tsinghua University, Beijing, 100084, China
[b] Division of Information Science, Tsinghua Shenzhen International Graduate School, Shenzhen, 518055, China
[c] Inspur Yunzhou Industrial Internet Co., Ltd, Jinan, 250101, China

ARTICLE INFO

Keywords:
Industrial Anomaly Detection
Unsupervised learning
Multimodal learning
Feature mapping
RGB-3D fusion

ABSTRACT

While RGB-based methods have been extensively studied in Industrial Anomaly Detection (IAD), effectively incorporating point cloud data remains challenging. Alongside prevalent memory bank-based approaches, recent research has explored cross-modal feature mapping for multimodal IAD, achieving notable performance and efficient inference. However, cross-modal feature mapping, while effective for detecting anomalies in feature correspondences, struggles to identify those exclusive to a single modality, due to the inherent one-to-many mapping between 2D and 3D data. To overcome this limitation, we propose **Cross-modal Prediction and Intra-modal Reconstruction (CPIR)**, a novel multimodal anomaly detection method. First, we introduce a **Bidirectional Feature Mapping (BFM)** framework that integrates intra-modal reconstruction tasks with cross-modal prediction tasks, enhancing single-modality anomaly detection while maintaining effective cross-modal consistency learning. Second, we propose a novel network architecture, **Latent Bridged Modal Mapping Module (LB3M)**, which introduces a shared latent intermediate state to decouple feature mapping across modalities into mappings between each modality and a shared intermediate state. This design was initially proposed to effectively complete prediction and reconstruction tasks with minimal parameters. However, it also enabled the network to learn more comprehensive feature patterns, significantly improving anomaly detection capabilities. Experiments on the MVTec 3D-AD dataset demonstrate that CPIR outperforms state-of-the-art methods in both anomaly detection and segmentation tasks, while excelling in few-shot learning scenarios.

1. Introduction

Industrial anomaly detection aims to identify anomalies and defects in industrial products, which is critical for maintaining production quality and efficiency. Due to the scarcity of anomalous samples in real-world industrial scenarios, research in this field primarily focuses on unsupervised approaches. While substantial progress has been made using RGB data as input, 3D data inherently excels in detecting volumetric and structural anomalies. Therefore, exploring multimodal anomaly detection by integrating 3D and RGB data is both meaningful and promising.

For multimodal Industrial Anomaly Detection (IAD), existing methods typically follow the embedding-based paradigm, where pre-trained models are used for multimodal feature extraction, followed by anomaly feature detection. Extensive research has investigated the selection and application technique of pre-trained feature extractors, resulting in several well-established approaches [1–3]. Building on these mainstream feature extraction protocols, our work focuses on the subsequent challenge of designing effective anomaly detection algorithms. Most existing methods rely on memory banks [1,2,4,5], which construct memory banks to represent normal feature distributions. However, these methods suffer from notable limitations [3]: they are memory-inefficient, computationally expensive during inference, and heavily dependent on a large number of normal samples to construct representative memory banks. These drawbacks severely limit their applicability in industrial scenarios, particularly in few-shot settings.

A recent study [3] proposes cross-modal feature prediction for multimodal IAD, where prediction discrepancies indicate abnormality levels. The model, composed of shallow MLPs, achieves comparable performance to memory bank-based methods while requiring few parameters.

* Corresponding author.
E-mail address: hbq@tsinghua.edu.cn (B. Huang).
[1] Wen Shangguan and Hongqiang Wu contributed equally to this work.

https://doi.org/10.1016/j.aei.2025.103240
Received 5 December 2024; Received in revised form 13 February 2025; Accepted 28 February 2025
Available online 14 March 2025

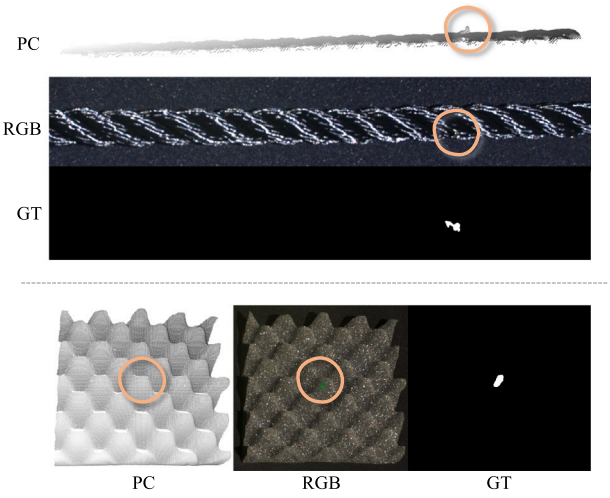


Fig. 1. Examples of single-modality anomalies. Top: Minor contamination anomalies on rope products are barely visible in the RGB modality but exhibit significant feature differences in the 3D modality. Bottom: Color anomalies on foam products are indistinguishable in the point cloud modality but are highly noticeable in the RGB modality.

However, while cross-modal feature prediction methods effectively learn feature correspondence patterns between modalities for efficient anomaly detection, they inherently struggle to identify anomalies present only in a single modality. Fig. 1 illustrates two typical single-modality anomalies: minor contamination anomalies and color anomalies. For instance, in the upper part of Fig. 1, minor contamination anomalies on rope products are barely visible in the RGB modality but exhibit significant feature differences in the 3D modality. Similarly, in the lower part of Fig. 1, color anomalies on foam products are indistinguishable in the point cloud modality but are highly noticeable in the RGB modality.

This asymmetry between modalities leads to a one-to-many mapping, where similar features in one modality may correspond to entirely different features in another. During training, such one-to-many mappings may also exist in normal samples. For example, variations in lighting or shadows caused by different camera angles can significantly alter 2D modality features while leaving 3D modality features unchanged. In these cases, the model struggles to accurately predict feature mappings across modalities, leading to high prediction deviations even for normal samples. These deviations can be confused with actual anomalous regions during inference, resulting in higher false positive rates and reduced detection performance.

To overcome this issue, we propose a novel multimodal anomaly detection method based on feature mapping, named CPIR. Specifically, we introduce a Bidirectional Feature Mapping Framework (BFM) that combines intra-modal reconstruction with cross-modal prediction, enabling efficient detection across various multimodal anomaly scenarios. The intra-modal reconstruction task provides the model with a single-modal detection perspective, allowing it to learn unique modality-specific knowledge while cross-modal prediction focuses on learning the correspondence between multimodal features. For anomalies that primarily exist within a single modality, when cross-modal predictions fail, the intra-modal reconstruction results can clearly indicate the abnormal patterns.

To implement this framework, we design a novel network structure that efficiently handles feature mappings across modalities and within individual modalities. Specifically, we propose a Latent Bridged Modal Mapping Module (LB3M). By introducing a latent intermediate feature space that bridges the 3D and 2D feature spaces, any feature mapping between modalities can be decomposed into two sub-processes. The first sub-process maps features from the source feature space to the intermediate space, serving as an encoding step. The second sub-process maps features from the intermediate space to the target feature space, functioning as a decoding step. Consequently, for both cross-modal and intra-modal feature mappings, the model only needs to learn the transformation processes between the intermediate space and each modality-specific feature space. This design significantly reduces the number of parameters required, improving both efficiency and generalization. Such a multi-task learning strategy provides the model with additional perspectives, helping it acquire more comprehensive feature patterns.

In summary, our main contributions are as follows:

1. We propose a novel multimodal anomaly detection method, CPIR, which outperforms state-of-the-art methods on the MVTec 3D-AD [6] dataset, achieving superior results in anomaly classification and segmentation tasks, as well as in few-shot learning settings.
2. We introduce a new unsupervised multimodal anomaly detection framework, Bidirectional Feature Mapping (BFM), which combines cross-modal prediction and intra-modal reconstruction, enabling robust and accurate anomaly detection through comprehensive feature modeling.
3. We design a novel prediction and reconstruction network architecture, LB3M, based on a latent intermediate representation, which enables efficient parameter sharing across modalities and tasks, allowing the feature mapping module to learn more diverse and representative feature patterns.

2. Related works

Industrial Anomaly Detection (IAD) has traditionally focused on 2D-based methods, which have received significant attention and development within the research community. Consequently, algorithms for multimodal industrial anomaly detection using RGB-D data are largely inspired by approaches developed for 2D IAD. In this section, we first review the prevailing paradigms in 2D IAD and then discuss the solutions proposed for multimodal 3D IAD.

2.1. Unsupervised image anomaly detection

The research on 2D IAD has primarily been driven by the MVTec AD benchmark [7], where the task involves training solely on normal samples and testing on a mixture of normal and abnormal samples. The goal is to enable models trained on normal data to effectively identify anomalies in the test set. Current methods can be broadly categorized into two groups [8]: reconstruction-based methods [9–17] and feature embedding-based methods [18–21].

Reconstruction-based methods. Reconstruction-based methods are predicated on the assumption that models trained exclusively on normal samples exhibit significantly diminished reconstruction capabilities when encountering anomalous regions. Anomalies are therefore identified by comparing the reconstruction output to the input, with significant discrepancies indicating abnormal areas. These methods focus on the design of the reconstruction module, which can take various forms depending on the underlying generative model, such as autoencoders [9,10], generative adversarial networks (GANs) [11–13], transformers [14,15], or diffusion models [16,17].

Feature embedding-based methods. Feature embedding-based methods leverage pretrained feature extractors to directly utilize high-level features. These methods aim to learn the distribution of normal features and identify deviations that signify anomalies. The pipeline typically involves two stages: feature extraction and anomaly detection. ResNet [18,22,23] and Vision Transformer (ViT) [24,25] are widely used for feature extraction in 2D IAD. For anomaly detection, three main approaches are employed: distribution-based methods, memory bank-based methods, and feature reconstruction-based methods.

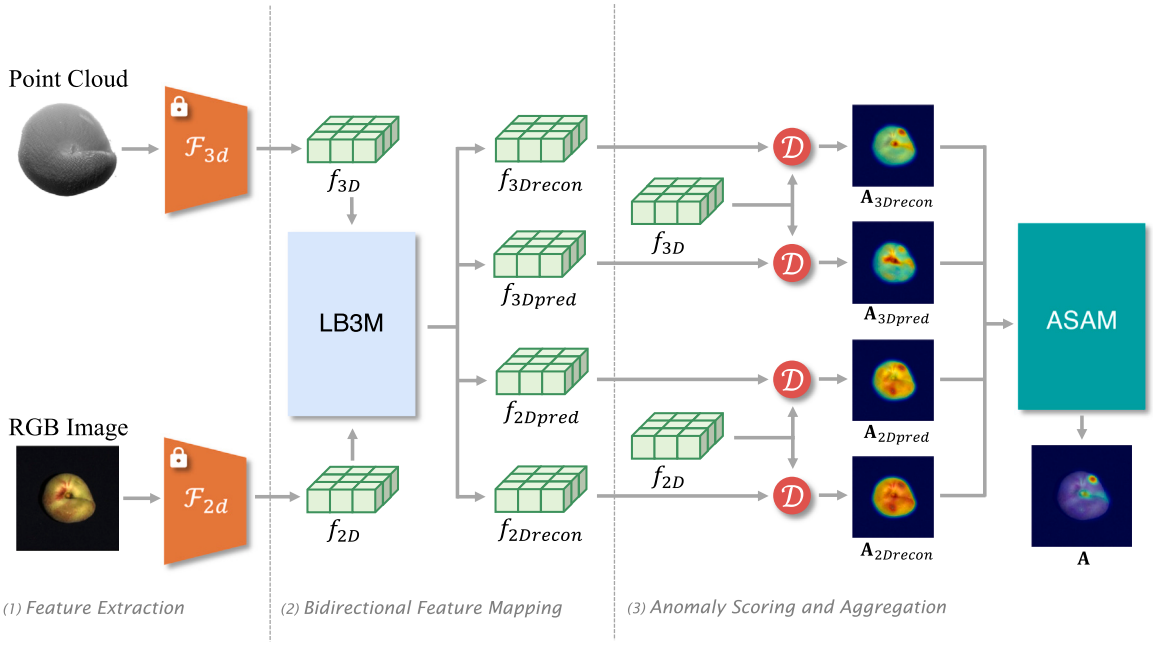


Fig. 2. The pipeline of CPIR. First, a pre-trained feature extractor is used to extract features from the point cloud and RGB modality respectively. Then, the LB3M module calculates the results of mutual prediction between modes and reconstruction within modes. Finally, the aggregation module compares the results with the original features and aggregates the results.

Distribution-based methods model the distribution of normal features using techniques such as multivariate Gaussian fitting [26] or normalizing flows [21]. These models compute the anomaly probability of test samples by measuring deviations from the learned distribution.

Memory bank-based methods [18,27–29] store a memory of normal features and use matching techniques during inference to identify anomalies. For instance, SPADE [27] employs KNN for feature matching and L2 distance for anomaly scoring. Subsequent works improve upon this by refining memory matching [28,29] or distance calculation [18].

Feature reconstruction-based methods [14,30–32] focus on reconstructing features rather than input images. These methods assume that abnormal features cannot be accurately reconstructed by a model trained on normal data. For example, teacher–student architectures [19,20] are used for feature reconstruction, while diffusion models and feature editing are explored in [32].

Each approach has its strengths and limitations. Feature embedding-based methods benefit from the superior performance of pre-trained feature extractors [33], but suffer from domain gaps between the pre-trained dataset and IAD tasks. In contrast, reconstruction-based methods rely less on pre-trained models and leverage self-supervision, making them more robust to domain shifts. However, their dependence on generative models can pose challenges in data-scarce or complex scenarios, where reliable reconstruction is difficult to achieve.

2.2. Multimodal RGB-3D anomaly detection

Industrial Anomaly Detection (IAD) in multimodal RGB-3D settings has been extensively studied using the MVTec 3D-AD dataset [6], which comprises aligned RGB images and corresponding depth maps captured through structured light sensing. Leveraging this dataset with 3D geometric information, researchers have explored both 3D-specific anomaly detection approaches and multimodal methods that jointly utilize RGB and 3D data. In the following sections, we first briefly review approaches focusing solely on 3D data, followed by a detailed discussion of methods that integrate RGB and 3D modalities for comprehensive anomaly detection.

3D-only Anomaly Detection Methods. Early explorations in 3D anomaly detection have yielded several notable approaches. 3D-ST [34] proposes a task-specific local descriptor trained through self-supervised reconstruction, employing a teacher–student paradigm for anomaly detection. CPMF [35] innovatively bridges the gap between 3D and 2D domains by projecting point clouds into multi-view images, leveraging pre-trained image feature extractors to complement hand-crafted 3D features. With the introduction of high-fidelity 3D anomaly detection datasets like Real3D-AD [36] and Anomaly-ShapeNet [37], more sophisticated approaches have emerged. R3D-AD [38] leverages diffusion models for point cloud reconstruction and introduces a novel PatchGen strategy for realistic anomaly simulation. ISMP [39] introduces an innovative Internal Spatial Modality Perception framework that explores rich internal structural information within 3D samples, achieving superior detection performance.

While these 3D-only methods primarily focus on developing more effective and efficient geometric feature extraction strategies, multimodal approaches emphasize the effective fusion and utilization of features from multiple modalities. For multimodal IAD, essential data preprocessing techniques and two primary methodological paradigms will be discussed. For the predominant feature-based paradigm, we provide a detailed examination from both feature extraction and anomaly detection perspectives.

Data Preprocessing. For multimodal IAD, 3D data often requires significant preprocessing to address challenges such as background noise, outliers, and point cloud sparsity. BTF [4] introduced a pipeline that first removes the background plane using RANSAC, followed by outlier removal via DB-SCAN. This preprocessing step has since been widely adopted in subsequent works. Addressing the inherent sparsity of point clouds caused by varying acquisition angles and quality, Li et al. [40] proposed a point cloud upsampling technique specifically designed for industrial 3D data, demonstrating improved performance in downstream anomaly detection tasks.

Methodological Paradigms. After preprocessing, multimodal RGB-3D IAD methods can be broadly categorized into reconstruction-based and feature-based methods.

Due to the relative immaturity of 3D reconstruction techniques compared to image reconstruction, reconstruction-based methods face

greater challenges in unsupervised anomaly detection settings and are therefore less prevalent. Nonetheless, some methods have achieved promising results. For instance, 3DSR [41] and EasyNet [42] generate pseudo-anomalous inputs akin to DRAEM [43] and train models to reconstruct their corresponding normal samples in a self-supervised manner.

In contrast, the feature-based paradigm has emerged as the dominant approach for multimodal RGB-3D IAD. This paradigm can be further divided into two components: feature extraction and anomaly feature detection.

Feature Extraction. Feature extraction in multimodal IAD aims to capture informative representations for both RGB and 3D data. While RGB feature extractors are typically pre-trained on large-scale image datasets, 3D feature extraction poses unique challenges due to the distinct characteristics of industrial 3D data in MVTec 3D-AD, which differs from large-scale 3D datasets like ScanNet [44].

Early works such as BTF [4] utilized handcrafted 3D descriptors (e.g., FPFH [45]) for anomaly detection. Later, M3DM [1] introduced pre-trained models like PointMAE [46] to extract 3D features, projecting them onto the 2D plane based on the one-to-one correspondence between 3D points and 2D pixels. This pipeline has become the standard approach in subsequent research [2,3,40]. Building on this, CFM [3] demonstrated that selectively pruning transformer layers in pre-trained models improves computational efficiency without degrading feature quality. In parallel, Shape-guided [5] leveraged PointNet [47] with NFIDF [48] to extract 3D features, achieving competitive performance.

While our primary contribution lies in advancing anomaly feature detection, we build upon the widely adopted feature extraction pipeline [1–3,40], which has become a standard in recent works, to ensure fair comparisons.

Anomaly Feature Detection. The anomaly feature detection stage focuses on identifying abnormal patterns within the extracted features. In multimodal RGB-3D IAD, this process involves modeling both intra-modal characteristics and cross-modal correspondences. The dominant approach in current research relies on memory bank-based methods to address this challenge.

M3DM [1] fused RGB and 3D features via contrastive learning and applied PatchCore [18] independently to each modality and their fused representations. Anomaly scores were then aggregated using OCSVM [49]. Shape-guided [5] introduced memory bank matching to model 2D-3D correspondences, refining 2D matching results using 3D memory banks. ITNM [50] employed weighted feature concatenation for cross-modal fusion and proposed a template neighborhood matching strategy to improve anomaly detection. LSFA [2] enhanced M3DM by introducing global alignment mechanisms and improving feature compactness, achieving better detection performance.

Despite their effectiveness, memory bank-based methods face several limitations. First, they require sampling a coreset from the normal sample set to represent the normal distribution, which often results in low memory utilization. Additionally, as inference heavily relies on querying the memory bank, this process is difficult to parallelize and typically leads to low runtime efficiency.

Recently, CFM [3] proposed an alternative paradigm resembling feature reconstruction. It adopted a lightweight MLP to predict features from one modality to another, enabling cross-modal mappings. By training on the cross-modal mapping task, the model can infer anomaly scores based on the deviation of predicted features from the target modality. This approach eliminates the spatial and temporal inefficiencies of memory bank-based methods while achieving considerable results. However, as discussed in Section 1, it struggles to detect anomalies confined to a single modality.

To address this limitation, we aim to expand the anomaly detection capability of the feature mapping framework by introducing an intra-modal reconstruction task and exploring a more effective mapping module architecture.

3. Method

3.1. Overview

In this section, we present the overall framework of the proposed CPIR method. As illustrated in Fig. 2, the model is composed of three main components: (1) **Feature Extraction**, (2) **Bidirectional Feature Mapping (BFM)**, which is implemented by **Latent Bridged Modal Mapping Module (LB3M)**, and (3) **Anomaly Scoring and Aggregation Module (ASAM)**.

The CPIR pipeline takes a 3D point cloud and a 2D image as input and generates anomaly score maps $\mathbf{A}_{agg}$ as output. First, during the **Feature Extraction** stage, pre-trained models are utilized to extract modality-specific features, resulting in feature representations f_{3D} and f_{2D} for the 3D and 2D inputs, respectively. These features serve as the foundation for subsequent processing.

Next, the core of CPIR lies in **BFM**, a framework that defines two complementary tasks: (1) Cross-Modal Prediction, which captures the consistency and correspondence between different modalities, and (2) Intra-Modal Reconstruction, which focuses on learning modality-specific feature patterns.

These two tasks in the BFM framework are jointly implemented through the proposed **LB3M**, which generates predictions and reconstructions (f_{3Dpred}, f_{2Dpred}, $f_{3Drecon}$, $f_{2Drecon}$). During training, these outputs are compared with the original input features to compute the training loss. During inference, a distance metric D is applied to these outputs to compute the initial anomaly scores.

Finally, in **ASAM**, the anomaly scores derived from multiple tasks and modalities are combined using the proposed module, producing the final anomaly score maps, $\mathbf{A}_{agg}$.

The remainder of this section is organized as follows: Section 3.2 describes the Feature Extraction process, Section 3.3 introduces the BFM framework, Section 3.4 details the design of the LB3M module, and Section 3.5 presents the ASAM module.

3.2. Feature extraction

Our method operates on extracted feature maps and is therefore agnostic to the choice of feature extraction scheme. For fair comparison, we adopt a widely-used feature extraction pipeline that has proven effective in multimodal IAD tasks. Let (I_i, P_i) denote the ith input sample, where $I_i \in \mathbb{R}^{H\times W\times 3}$ represents the RGB image, and $P_i \in \mathbb{R}^{N\times 3}$ denotes the point cloud.

2D Feature Extraction. For 2D feature extraction, we employ DINO, a vision transformer-based pre-trained model, as the feature extractor Φ_{2D}. The input image is divided into patches of size p, producing feature representations:

$$\phi_i^{2d} = \Phi_{2d}(I_i) \in \mathbb{R}^{\frac{H}{p}\times\frac{W}{p}\times C_{2d}}. \tag{1}$$

To ensure spatial correspondence with multimodal features, these features are bilinearly upsampled back to the original image resolution:

$$f_{2d}^i = \mathcal{F}_{rgb}(I_i) = g_{bilinear}\left(\phi_i^{2d}\right) \in \mathbb{R}^{H\times W\times C_{2d}}. \tag{2}$$

3D Feature Extraction. For 3D feature extraction, we employ PointMAE, a PointTransformer-based pretrained model, as the feature extractor Φ_{3d}. Following common practice, the point cloud is first downsampled using farthest point sampling (FPS) to obtain M groups, each containing S points. The extracted 3D features are represented as:

$$\phi_i^{3d} = \Phi_{3d}(P_i) \in \mathbb{R}^{M\times C_{3d}}. \tag{3}$$

Since the FPS-selected center points are spatially non-uniform, we apply feature interpolation following [51] to restore the original point cloud resolution:

$$f_{3d}^i = \mathcal{F}_{pc}(P_i) = g_{PFI}\left(\phi_i^{3d}\right) \in \mathbb{R}^{N\times C_{3d}}. \tag{4}$$

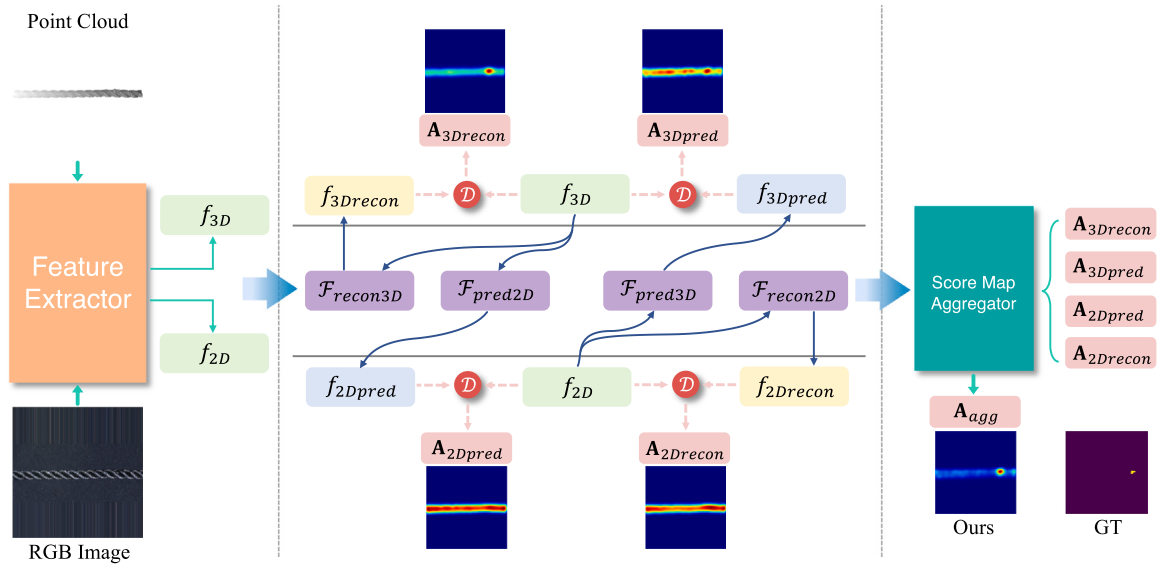


Fig. 3. Illustration of the Bidirectional Feature Mapping Framework (BFM) for multimodal Industrial Anomaly Detection (IAD). The central part of the figure illustrates the bidirectional mapping process, where the two input features, f_{3D} and f_{2D}, are transformed through cross-modal predictions and intra-modal reconstructions via four trained mapping functions: $\mathcal{F}_{pred3D}$, $\mathcal{F}_{pred2D}$, $\mathcal{F}_{recon3D}$, and $\mathcal{F}_{recon2D}$. The outputs of these mappings are compared with the corresponding input features using a distance metric D to compute similarity scores, which are visualized as four anomaly sub-score maps. Finally, a Score Map Aggregator combines these sub-score maps into a unified anomaly score map.

For the MVTec 3D-AD dataset [6], the point cloud data is acquired using an RGB-D camera, where each 3D point naturally corresponds to a pixel in the RGB image. As a result, the total number of points is $N = H \times W$.

Pruning for Efficiency. Notably, we apply pruning to both 2D and 3D feature extraction transformers. Based on the observations in [3], using only shallow-layer features of transformers can significantly improve computational efficiency and reduce the number of parameters, with minimal performance degradation. Therefore, we utilize the feature outputs from only the first l layers of both models.

3.3. BFM: Bidirectional feature mapping framework

Although CFM [3] first introduced cross-modal prediction for multimodal IAD, as discussed in Section 1, this cross-modal prediction paradigm suffers from limitations when anomalies exist solely in one modality. To address this limitation and better model multimodal feature patterns, we propose a novel **Bidirectional Feature Mapping (BFM)** framework, which integrates intra-modal reconstruction and cross-modal prediction tasks for unsupervised multimodal IAD.

As illustrated in Fig. 3, the core module (central part of the figure) takes multimodal features $f_{3D} \in \mathbb{R}^{H\times W\times C_{3D}}$ and $f_{2D} \in \mathbb{R}^{H\times W\times C_{2D}}$ as inputs and outputs four mapping results: f_{3Dpred} and f_{2Dpred} for cross-modal prediction, and $f_{3Drecon}$ and $f_{2Drecon}$ for intra-modal reconstruction. The mapping functions $\mathcal{F}_{pred3D}$ and $\mathcal{F}_{pred2D}$ perform cross-modal prediction between modalities, while $\mathcal{F}_{recon2D}$ and $\mathcal{F}_{recon3D}$ handle intra-modal reconstruction within each modality.

The feature mapping processes can be formalized as follows. For cross-modal prediction, the goal is to map features from one modality to another by minimizing a similarity distance metric D. For example, when mapping 2D features to 3D features, the task can be expressed as:

$$\mathcal{F}_{pred3D} = \arg\min_{\mathcal{F}} D\left[\mathcal{F}(f_{2D}), f_{3D}\right]. \tag{5}$$

Similarly, $\mathcal{F}_{pred2D}$, $\mathcal{F}_{recon2D}$, and $\mathcal{F}_{recon3D}$ can be defined using the corresponding distance metrics. By training these feature mapping tasks, the framework aims to model the normal feature distributions across and within modalities. This enables the mapping results (e.g., f_{3Dpred}) to approximate the corresponding modality features of normal samples (e.g., f_{3D}). When the similarity between the mapping result and the observed features (e.g., f_{3Dpred} vs. f_{3D}) is low, the sample is more likely to belong to an anomaly. Specifically, for anomaly detection, the anomaly score measures the likelihood of a sample being anomalous and can be calculated as:

$$\mathbf{A}_{3Dpred} = D(f_{3Dpred}, f_{3D}). \tag{6}$$

where $\mathbf{A}_{3Dpred}$ denotes the anomaly score when predicting 3D features from 2D features. Similarly, anomaly scores can be computed for $\mathbf{A}_{2Dpred}$, $\mathbf{A}_{3Drecon}$ and $\mathbf{A}_{2Drecon}$.

In this way, cross-modal prediction captures the correspondence between different modalities, enabling the model to detect anomalies that emerge from inconsistencies across modalities. Intra-modal reconstruction captures fine-grained patterns within each modality, effectively identifying anomalies confined to a single modality. These complementary tasks ensure robust and comprehensive anomaly detection.

3.4. LB3M: Latent bridged modal mapping module

To achieve both cross-modal prediction and intra-modal reconstruction tasks within a unified network architecture, we propose the **Latent Bridged Cross-modal Prediction and Intra-modal Reconstruction Module (LB3M)**. As shown in Fig. 4, the core innovation of LB3M is the introduction of an latent feature space between the 3D and 2D feature spaces. With this design, any feature mapping between modalities can be decomposed into two sub-processes: first, mapping features from the source feature space to the latent space, and then mapping from the latent space to the target feature space. This decomposition not only simplifies the complex feature mapping process but also provides a unified intermediate representation for different mapping tasks.

The architecture of LB3M consists of four basic feature mapping modules: $\mathcal{M}_{2D\to mid}$ and $\mathcal{M}_{3D\to mid}$ map features from their respective modalities to the latent feature space, while $\mathcal{M}_{mid\to 2D}$ and $\mathcal{M}_{mid\to 3D}$ map latent features back to the corresponding modality feature spaces. By combining these modules, LB3M achieves both cross-modal prediction and intra-modal reconstruction, as formalized below:

$$\begin{aligned}
\mathcal{F}_{pred3D}(\cdot) &= \mathcal{M}_{mid\to 3D}(\mathcal{M}_{2D\to mid}(\cdot)),\\
\mathcal{F}_{pred2D}(\cdot) &= \mathcal{M}_{mid\to 2D}(\mathcal{M}_{3D\to mid}(\cdot)),\\
\mathcal{F}_{recon3D}(\cdot) &= \mathcal{M}_{mid\to 3D}(\mathcal{M}_{3D\to mid}(\cdot)),\\
\mathcal{F}_{recon2D}(\cdot) &= \mathcal{M}_{mid\to 2D}(\mathcal{M}_{2D\to mid}(\cdot)).
\end{aligned} \tag{7}$$

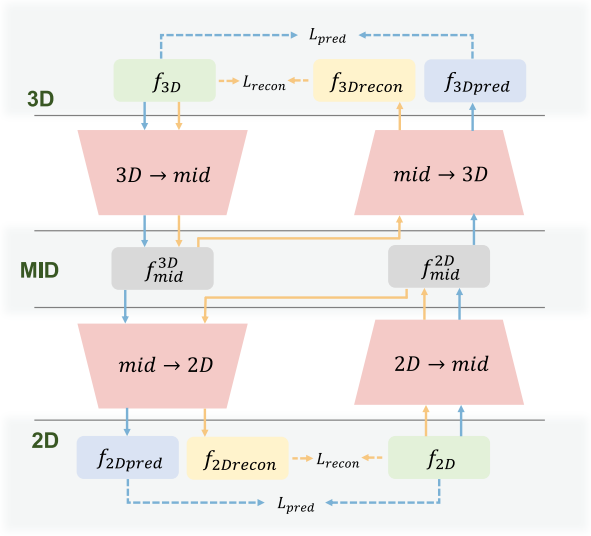


Fig. 4. Illustration of the Latent Bridged Modal Mapping Module (LB3M). The LB3M consists of modules that model the mapping from the middle latent space to each modality (3D and 2D) and vice versa. In the figure, yellow solid arrows represent the feature flow for intra-modal reconstruction, while blue solid arrows indicate the feature flow for cross-modal prediction. During training, the mapped results are compared with the original features to compute the loss.

To optimize the performance of the model, we define two loss functions: the prediction loss L_{pred} and the reconstruction loss L_{recon}, which measure the accuracy of cross-modal feature prediction and the fidelity of intra-modal feature reconstruction, respectively. The overall loss function L is a weighted sum of the two:

$$L_{pred} = \mathcal{D}(f_{3D}, f_{3Dpred}) + \mathcal{D}(f_{2D}, f_{2Dpred}), \tag{8}$$

$$L_{recon} = \mathcal{D}(f_{3D}, f_{3Drecon}) + \mathcal{D}(f_{2D}, f_{2Drecon}), \tag{9}$$

$$L = L_{pred} + \beta L_{recon}. \tag{10}$$

From a structural perspective, LB3M can be interpreted as four interlinked autoencoders with shared components. In this framework, the encoders map features from the source modality to the latent space, while the decoders map latent features back to the target modality space. Each modality's prediction and reconstruction tasks share the same encoder, enabling the encoder to learn feature representations of the same object under different tasks. Similarly, each decoder is responsible for both prediction and reconstruction, requiring it to accurately decode latent features into the target modality.

This shared design allows the model to learn more comprehensive representations of the target modality's feature patterns. Supervised by both prediction and reconstruction tasks, the shared encoder and decoder are able to transfer and reuse knowledge across tasks, enabling all modules to learn robust and general latent feature representations. These representations are required to preserve the critical information of the original modality in order to facilitate knowledge sharing between prediction and reconstruction tasks, ultimately enhancing the model's capacity to generalize effectively across diverse modalities and tasks.

Moreover, previous studies have shown that shallow MLPs are effective enough for modeling cross-modal feature mapping [3]. Inspired by this, we adopt shallow MLPs as the specific structure for the mapping modules $\mathcal{M}$. This choice allows LB3M to maintain high inference efficiency while minimizing the number of parameters.

3.5. ASAM: Anomaly scoring and aggregation module

After obtaining the prediction and reconstruction results from the LB3M module, we first compute the corresponding four anomaly score maps based on Eq. (6). Each anomaly score map represents the likelihood of pixels being anomalous from the perspective of a single feature mapping task.

When designing the anomaly score aggregation strategy, we consider the challenges in multimodal anomaly detection scenarios: different products and anomaly types often exhibit varying patterns across different feature mapping tasks. To address this, we propose using a multiplicative logic AND operation for anomaly score aggregation. This is motivated by the observation that when an anomaly is not prominent in a specific task or modality, its anomaly score may be confused with challenging normal patterns, leading to high false-positive rates. The logic AND operation helps mitigate this issue by filtering out anomaly scores that are only significant in individual tasks, while retaining regions that are consistently identified as anomalous across all tasks. Mathematically, the aggregated anomaly score is computed as the geometric mean of the four individual score maps:

$$\mathbf{A}_{agg} = \sqrt{\mathbf{A}_{3Drecon} \cdot \mathbf{A}_{2Dpred} \cdot \mathbf{A}_{2Drecon} \cdot \mathbf{A}_{3Dpred}}, \tag{11}$$

Finally, following common practice [1,18], we apply Gaussian smoothing with $\sigma = 4$ to the final anomaly score map. The maximum anomaly score among all pixels is then used as the image-level anomaly score for the sample.

4. Experiments

4.1. Experiment settings

Datasets. We evaluate our method on the widely-used MVTec-3D AD [6] dataset and the Eycandies [52] dataset. The MVTec-3D AD dataset comprises 10 distinct categories, with a total of 2656 training samples and 1137 test samples. Each category contains four to five anomaly subcategories. All samples are captured using a 3D structured light camera, resulting in monocular RGB-D images consisting of an RGB image and a corresponding depth map. While the depth map provides high-precision geometric information analogous to a point cloud, it captures objects from a single viewpoint, leading to incomplete spatial coverage and missing occluded regions.

The Eycandies dataset is a synthetic dataset specifically designed for multimodal anomaly detection, featuring RGB images along with depth and surface normal maps captured under various lighting conditions. The dataset contains 1500 samples per category, divided into 1000 training, 100 validation, and 400 test samples. While training and validation sets contain only normal samples, the test set is evenly split between normal and anomalous samples. Each category includes ten distinct anomaly types, with the anomalous test samples carefully balanced: 40 samples for each of the four basic anomaly types, plus 40 samples containing combined anomalies. Notably, all anomalies are procedurally generated with precise ground-truth annotations, ensuring high-quality labels without manual intervention.

Evaluation Metrics. We adopt the comprehensive evaluation framework established by [4], which includes both image-level and pixel-level metrics. For image-level evaluation, we use the Image-level Area Under the Receiver Operating Characteristic curve (I-AUROC). For pixel-level evaluation, we employ two metrics: the Pixel-level Area Under the Receiver Operating Characteristic curve (P-AUROC) and the Per-Region Overlap (PRO) metric. The PRO metric [4] measures the mean relative overlap between predicted anomalous regions and ground truth connected components. Following the protocol of [3], we report PRO performance under two false positive rate (FPR) thresholds: AUPRO@30% and AUPRO@1%, corresponding to FPR thresholds of 0.3 and 0.01, respectively. For all metrics, higher values indicate better performance.

Table 1
Anomaly detection and localization performance on MVTec 3D-AD dataset [6]. Each cell contains paired scores (I-AUROC, AUPRO@30%). The best and second-best results for each category are highlighted in red and green, respectively.

Category	BTF [4] (CVPR2023)	M3DM [1] (CVPR2023)	Shape-Guided [5] (ICML2023)	CFM [3] (CVPR2024)	LSFA [2] (ECCV2024)	DAUP [40] (ADVEI2024)	Ours
Bagel	93.8 / 97.6	99.4 / 97.0	98.6 / 98.1	99.4 / 97.9	100 / 98.6	99.6 / 97.6	99.9 / 98.1
Cable Gland	76.5 / 96.7	90.9 / 97.1	89.4 / 97.3	88.8 / 97.2	93.9 / 97.4	88.9 / 97.7	88.5 / 97.3
Carrot	97.2 / 97.9	97.2 / 97.9	98.3 / 98.2	98.4 / 98.2	98.2 / 98.1	99.6 / 98.0	97.6 / 98.2
Cookie	88.8 / 97.4	97.6 / 95.0	99.1 / 97.1	99.3 / 94.5	98.9 / 94.6	99.8 / 96.0	99.9 / 96.8
Dowel	96.0 / 97.1	96.0 / 94.1	97.6 / 96.2	98.0 / 95.0	96.1 / 92.5	97.7 / 92.4	100 / 97.3
Foam	66.4 / 88.4	94.2 / 93.2	85.7 / 97.8	88.8 / 96.8	95.1 / 94.1	93.9 / 96.6	91.6 / 97.5
Peach	90.4 / 97.6	97.3 / 97.7	99.0 / 98.1	94.1 / 98.0	98.3 / 98.3	98.3 / 98.1	99.5 / 98.3
Potato	92.9 / 98.1	89.9 / 97.1	96.5 / 98.3	94.3 / 98.2	96.2 / 98.3	98.6 / 97.8	98.1 / 98.3
Rope	98.2 / 95.9	97.2 / 97.1	96.0 / 97.4	98.0 / 97.5	98.9 / 97.4	97.9 / 97.2	99.8 / 98.2
Tire	72.6 / 97.1	85.0 / 97.5	86.9 / 97.5	95.3 / 98.1	95.1 / 98.3	96.0 / 98.0	99.2 / 98.3
Average	87.3 / 96.4	94.5 / 96.4	94.7 / 97.6	95.4 / 97.1	97.1 / 96.8	97.0 / 96.9	97.4 / 97.8

Table 2
Anomaly detection and localization performance on Eyecandies dataset [52]. Best results in **bold**.

Method	I-AUROC	P-AUROC	AUPRO@30%	AUPRO@1%
AST [53]	0.758	0.902	0.878	0.224
M3DM*[1]	0.881	**0.977**	0.886	0.331
CFM [3]	0.881	0.974	0.887	0.335
Ours	**0.886**	0.974	**0.894**	**0.346**

Preprocessing. For data preprocessing, we follow the standard protocol proposed in BTF [4], which addresses the critical challenge of background removal. Since objects are typically placed on a table during data acquisition, a significant portion of depth points corresponds to the background surface. To remove this background, we use the RANSAC algorithm [54] to estimate the dominant plane, which represents the background surface. Depth points within a distance threshold of 0.005 units from this plane are classified as background points and excluded from further processing. This preprocessing step ensures that subsequent feature extraction focuses exclusively on the object of interest, minimizing interference from irrelevant background information.

Implementation Details. For data preprocessing, all images are resized to a resolution of 224×224. For non-square images, the longer edge is resized to 224 while maintaining the original aspect ratio, and the remaining area is padded with zeros to form a square image.

For the backbone network, we use DINO ViT B/8 [24,55] and PointMAE [46] as the feature extractors. For DINO ViT B/8, only the first 8 layers of the transformer blocks are used, following the approach in CFM [3]. This modification has negligible impact on overall performance metrics while significantly accelerating model convergence. For PointMAE, we set the number of groups to 1024 and the group size to 128. After passing a $224 \times 224 \times 3$ input image through the backbone, it produces a $28 \times 28 \times 768$ 2D feature map, which is subsequently upsampled to a size of $224 \times 224 \times 768$. Similarly, a 224×224 point cloud is first processed into a 1024×1152 3D feature representation and then upsampled to a size of $224 \times 224 \times 1152$.

For the learnable modules, all four mapping modules are implemented as single hidden layer MLPs, where the number of hidden neurons is set to the average of the input and output feature sizes. The size of the latent feature space is set to 960, which is the average of the feature dimensions from the two modalities.

For training, we use the Adam optimizer with a learning rate of 0.001. The model is trained for a total of 150 epochs with a batch size of 2. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU.

4.2. Main results

Quantitative Results.

Table 1 presents the anomaly detection and localization performance of various methods on the MVTec 3D-AD dataset [6]. Each cell in the table contains two evaluation metrics: the image-level I-AUROC score and the pixel-level AUPRO@30% score, represented in the format "(I-AUROC, AUPRO@30%)". To visually highlight performance differences, the best and second-best results for each category are marked in red and green, respectively.

As shown in the table, our method significantly outperforms existing 3D anomaly detection methods on key evaluation metrics such as I-AUROC and AUPRO@30%. Compared to the baseline method CFM [3], our approach achieves substantial improvements of +2.0% in I-AUROC and +0.7% in AUPRO@30%, under the same feature extraction settings. In the category-wise analysis, our method excels in categories such as dowel (wooden pin), peach, potato, rope, and tire, consistently achieving the best results. Furthermore, in most subcategories, we achieve I-AUROC scores exceeding 99%. Compared to CFM, our method achieves notable improvements in nearly all subcategories, except for cable gland.

By comparing the performance characteristics of different methods, we observe a trade-off between image-level and pixel-level metrics in current research. Methods optimized for image-level metrics, such as LSFA [2] and DAUP [40], achieve outstanding I-AUROC scores (> 97.0%) but exhibit relatively weaker pixel-level performance (AUPRO < 97.0%). Conversely, methods focusing on pixel-level precision, such as CFM [3] and Shape-Guided [5], achieve high AUPRO@30% scores (> 97.1%) but show relatively lower image-level performance (I-AUROC < 95.4%). In contrast, our method successfully overcomes this trade-off, achieving superior performance on both levels with I-AUROC of 97.4% and AUPRO@30% of 97.8%.

It is worth noting that our method still exhibits certain limitations in the cable gland and foam categories. These categories present challenges not only for our method but also for the entire field of 3D anomaly detection. This difficulty may be attributed to the complex geometric structures and unique data acquisition environments associated with these objects, pointing to potential directions for future research.

Additionally, we evaluate our method on the Eycandies dataset, with results shown in Table 2. For fair comparison, we report the results of M3DM by retraining their model using the full dataset, rather than their originally reported results which were based on a subset of data and extensively searched memory bank sizes. Our approach achieves state-of-the-art performance across multiple metrics, including I-AUROC (88.6%), AUPRO@30% (89.4%), and AUPRO@1% (34.6%). These results, consistent with our findings on MVTec 3D-AD, further demonstrate the effectiveness and generalizability of our method across different industrial multimodal anomaly detection scenarios.

Qualitative Results. Fig. 5 illustrates the anomaly detection and segmentation results of CPIR on the MVTec 3D-AD dataset [6]. Compared to M3DM and CFM, CPIR significantly reduces false positives while maintaining a high detection rate through its dual feature mapping mechanisms, both cross-modal and intra-modal. Specifically, the

Table 3
Inference speed on MVTec-3D AD dataset.

Method	BTF [4]	AST [53]	M3DM [1]	CFM [3]	Ours
FPS	1.70	2.92	0.30	**12.29**	6.93
I-AUROC	0.865	0.937	0.945	0.954	**0.974**
AUPRO@30%	0.959	0.944	0.964	0.971	**0.978**

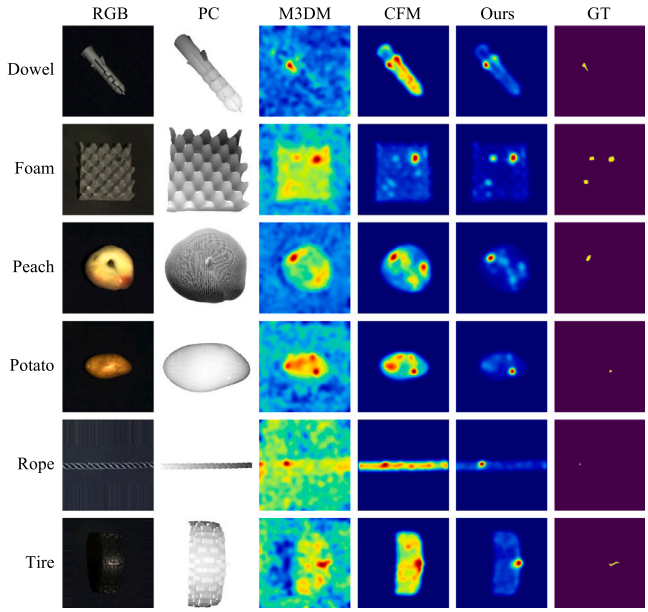


Fig. 5. Qualitative comparison on MVTec 3D-AD[6]. From left to right: the RGB image, a point cloud visualization, anomaly score maps generated by M3DM [1], CFM [3], and our method CPIR, followed by the ground truth annotations.

design of CPIR enables the model to better learn normal patterns across multiple modalities, effectively distinguishing between normal and anomalous regions. This capability not only ensures accurate detection of anomalous areas but also minimizes anomaly scores in normal regions. Such precise differentiation allows CPIR to achieve state-of-the-art performance on strict evaluation metrics such as AUPRO@1%. From the visualized anomaly score maps, CPIR demonstrates clearer boundaries and higher contrast, further validating its superiority in anomaly localization tasks.

Complexity Analysis. To evaluate computational efficiency, we compare our method with BTF [4], AST [53], M3DM [1], and CFM [3]. Our model extends CFM's architecture by incorporating two additional single-layer MLPs, which doubles the number of MLPs in the original CFM. As presented in Table 3, this modification reduces the processing speed of our method to approximately half that of CFM.

While this architectural enhancement introduces moderate computational overhead, our method still retains a significant speed advantage over other embedding-based approaches such as BTF, AST, and M3DM. This demonstrates the inherent efficiency of feature mapping-based methods. Furthermore, considering the trade-off between speed and performance, our method achieves state-of-the-art detection accuracy while maintaining competitive processing speeds, striking an effective balance between computational efficiency and detection performance.

4.3. Investigation on aggregation method.

Quantitative Results.

Table 4 compares three anomaly score aggregation methods on the MVTec 3D-AD dataset [6]: summation (Add), maximum value (Max), and our proposed multiplicative aggregation (Ours).

The maximum-based aggregation performs the worst across all metrics, with a significant drop in I-AUROC (0.906) and AUPRO@1% (0.410). This is because it amplifies the impact of false positives from noisy contributions in individual anomaly maps, making it unsuitable for robust multi-modal fusion.

In contrast, our multiplicative aggregation achieves the best performance across all metrics (e.g., I-AUROC 0.974, AUPRO@1% 0.474). By using multiplication to approximate a logical AND operation, this method effectively suppresses noise and emphasizes consistent anomaly signals, making it particularly well-suited for our BFM framework and its reliance on multi-modal feature fusion.

Qualitative Results. Fig. 6 visualizes the anomaly score maps and their aggregated results for CPIR. From top to bottom: color anomaly in foam, hole anomaly in peach, contamination anomaly in rope, and cut anomaly in tire. The results highlight the necessity of the intra-modal reconstruction task as a complement to the cross-modal prediction task. For single-modality anomalies (e.g., the first and third samples in Fig. 6), the branch that independently detects anomalies within the same modality achieves clearer and more accurate results by not relying on cross-modal information, which avoids the large number of false positives caused by the one-to-many mapping between modalities.

From the perspective of individual anomaly score maps, each map demonstrates specific strengths and weaknesses due to the complexity of the dataset. For example, $\mathbf{A}_{3Drecon}$ performs well in the first three samples but struggles to detect anomalies in the fourth sample. In contrast, $\mathbf{A}_{2Dpred}$ performs poorly in the first and third rows but plays a critical role in the second and fourth rows, particularly capturing anomalies in the fourth sample that other maps miss.

From the perspective of aggregation strategies, logical AND (approximated via multiplication) proves effective in reducing false positives when anomalies are only present in a single modality. For example, in the peach sample (second row), logical AND filters out false positives from individual maps by leveraging complementary detection mechanisms, isolating true anomalies and lowering overall false-positive rates.

4.4. Investigation on task paradigm and mapping module architecture

To evaluate the effectiveness of the LB3M architecture and the dual-task learning paradigm, we conducted a series of ablation experiments, as summarized in Table 5. The Architecture column lists the configurations of the mapping module, where MLP denotes using two MLP with single hidden layer as the structure of each $\mathcal{F}$ and LB3M denotes using our LB3M to implement each $\mathcal{F}$. L_{pred} and L_{recon} indicate whether the prediction and reconstruction tasks were included during training, respectively, while P.S. denotes the parameter size relative to a single MLP. CFM-Dual* employs a CFM-based structure with only two MLPs, where the intra-modal reconstruction task is approximated by performing cross-modal prediction twice, as illustrated in Fig. 7. To further explore the contributions of the parameter sharing mechanism in LB3M, Ours-Pred* and Ours-Recon* adopt the LB3M structure with dual-task training but utilize only the prediction or reconstruction branch during inference.

Task Paradigm: Cross-modal Feature Mapping vs. Bidirectional Feature Mapping.

The independent contribution of BFM is thoroughly evaluated using MLPs as the baseline mapping module. Results show that the prediction-only (CFM) and reconstruction-only (CFM-Recon) variants exhibit contrasting performances. The prediction task proves more effective in single-task scenarios, with CFM achieving an I-AUROC of 0.954 compared to 0.919 for CFM-Recon. As a concrete implementation of BFM, combining both tasks (CFM-Dual) results in significant performance gains, achieving an I-AUROC of 0.962 and AUPRO@30% of 0.977. This demonstrates the complementary nature of prediction and reconstruction tasks, where their synergy enables more comprehensive anomaly detection.

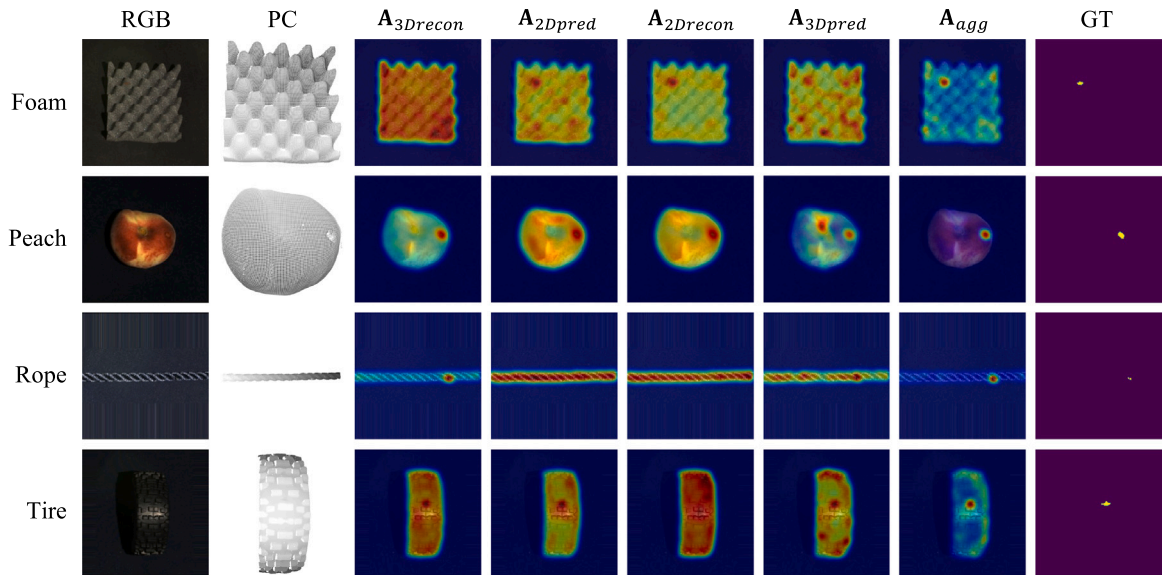


Fig. 6. Qualitative results of each anomaly score map and aggregation process. From left to right: the RGB image, point cloud visualization, $\mathbf{A}_{3Drecon}$, $\mathbf{A}_{2Dpred}$, $\mathbf{A}_{2Drecon}$, $\mathbf{A}_{3Dpred}$, the aggregated score map $\mathbf{A}_{agg}$, followed by the ground truth annotations.

Table 4
Investigation on the aggregation method. Best results in **bold**.

Method	Aggregate function	I-AUROC	P-AUROC	AUPRO@30%	AUPRO@1%
Add	$\mathbf{A}_{3Drecon} + \mathbf{A}_{2Dpred} + \mathbf{A}_{2Drecon} + \mathbf{A}_{3Dpred}$	0.972	0.995	0.974	0.467
Max	$\max\left(\mathbf{A}_{3Drecon}, \mathbf{A}_{3Dpred}, \mathbf{A}_{2Drecon}, \mathbf{A}_{2Dpred}\right)$	0.906	0.985	0.951	0.410
Ours	$\mathbf{A}_{3Drecon} \cdot \mathbf{A}_{2Dpred} \cdot \mathbf{A}_{2Drecon} \cdot \mathbf{A}_{3Dpred}$	**0.974**	**0.996**	**0.978**	**0.474**

Table 5
Investigation on task paradigm and mapping module architecture. Best results in **bold**.

Method	Architecture	L_{pred}	L_{recon}	I-AUROC	AUPRO@30%	AUPRO@1%	P.S.
CFM [3]	MLP	✓	✗	0.954	0.971	0.455	2×
CFM-Recon	MLP	✗	✓	0.919	0.968	0.436	2×
CFM-Dual*	MLP	✓	✓	0.955	0.971	0.453	2×
CFM-Dual	MLP	✓	✓	0.962	0.977	0.468	4×
Ours-Pred	LB3M	✓	✗	0.953	0.973	0.461	4×
Ours-Recon	LB3M	✗	✓	0.957	0.976	0.463	4×
Ours-Pred*	LB3M	✓	✓	0.957	0.974	0.461	4×
Ours-Recon*	LB3M	✓	✓	0.971	0.972	0.463	4×
Ours	LB3M	✓	✓	**0.974**	**0.978**	**0.474**	4×

Table 6
Investigation on parameter efficiency of different feature mapping architectures. Best results in **bold**.

Pred Arch.	Recon Arch.	L_{pred}	L_{recon}	I-AUROC	AUPRO@30%	AUPRO@1%	P.S.
MLP		✓	✗	0.954	0.971	0.455	2×
	MLP	✗	✓	0.919	0.968	0.436	2×
MLP	MLP	✓	✓	0.962	0.977	0.468	4×
	2MLP	✗	✓	0.957	0.976	0.463	4×
MLP	2MLP	✓	✓	0.968	**0.978**	0.472	6×
2MLP	2MLP	✓	✓	0.969	**0.978**	0.473	8×
LB3M	LB3M	✓	✓	**0.974**	**0.978**	**0.474**	4×

Table 7
Few-shot anomaly detection and segmentation on the MVTec 3D-AD [6] dataset. Best results in **bold**.

Method	I-AUROC				P-AUROC				AUPRO@30%				AUPRO@1%			
	5-shot	10-shot	50-shot	Full	5-shot	10-shot	50-shot	Full	5-shot	10-shot	50-shot	Full	5-shot	10-shot	50-shot	Full
BTF [4]	0.671	0.695	0.806	0.865	0.980	0.983	0.989	0.992	0.920	0.928	0.947	0.959	0.288	0.308	0.356	0.383
AST [53]	0.680	0.689	0.794	0.937	0.950	0.946	0.974	0.976	0.903	0.835	0.929	0.944	0.158	0.174	0.335	0.398
M3DM [1]	0.822	0.845	0.907	0.945	0.984	0.986	0.989	0.992	0.937	0.943	0.955	0.964	0.330	0.355	0.387	0.394
CFM [3]	0.811	0.845	0.906	0.954	0.986	0.987	0.991	0.993	0.949	0.954	0.965	0.971	0.382	0.398	0.431	0.455
Ours	**0.834**	**0.860**	**0.946**	**0.974**	**0.990**	**0.992**	**0.995**	**0.996**	**0.958**	**0.964**	**0.975**	**0.978**	**0.408**	**0.430**	**0.463**	**0.474**

Notably, the method that simulates reconstruction by performing consecutive cross-modal predictions (CFM-Dual*) does not lead to performance improvements (I-AUROC of only 0.955). This indicates that true bidirectional mapping, not simply stacking prediction tasks, is necessary for effective feature learning.

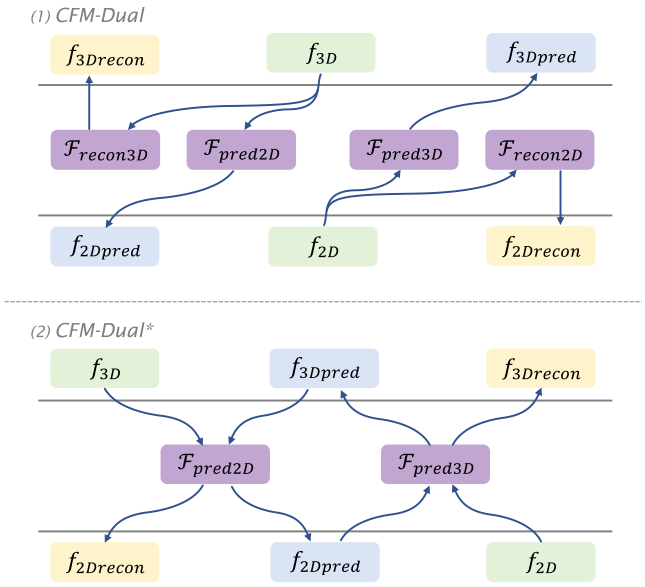


Fig. 7. Illustration of the differences between CFM-Dual and CFM-Dual*. Top: The structure of CFM-Dual, where both intra-modal reconstruction and cross-modal prediction tasks are performed using a single-layer MLP for each modality. Bottom: The structure of CFM-Dual*, where the intra-modal reconstruction task is approximated by sequentially applying the cross-modal prediction task twice.

Mapping Module Architecture: MLPs vs. LB3M.

The contribution of LB3M is thoroughly evaluated in two scenarios. In single-task settings, LB3M shows no significant improvements over MLPs. For the prediction task, Ours-Pred achieves an I-AUROC of 0.953, similar to CFM (0.954). For the reconstruction task, while Ours-Recon shows better performance than CFM-Recon (0.957 vs. 0.919), this gap mainly reflects the limited capacity of single-layer MLPs in handling reconstruction tasks, which can be observed by comparing rows 2 and 4 in Table 5. These results align with our design intention, as LB3M is specifically optimized for bidirectional feature mapping rather than single-task scenarios.

Under the BFM framework, for which LB3M is specifically designed, our complete model achieves state-of-the-art performance across all metrics (e.g., I-AUROC of 0.974, AUPRO@30% of 0.978). The substantial improvements over CFM-Dual highlight LB3M's significant contribution to bidirectional feature learning. Furthermore, we analyze the results of dual-task training with single-task inference to better understand its impact on feature learning. Ours-Pred* and Ours-Recon*, with only one branch being used during inference, both outperform their single-task counterparts. Notably, Ours-Recon* achieves an I-AUROC of 0.971, which matches the best image-level results of LSFA while outperforming it on pixel-level metrics. This strong performance demonstrates the benefits of LB3M's parameter-sharing mechanism among multiple encoders and decoders, enabling the model to learn more effective and comprehensive feature representations.

Analysis of Parameter Efficiency in Different Feature Mapping Architectures.

We investigate the parameter efficiency of different feature mapping architectures, with experimental results shown in Table 6. The Pred Arch. and Recon Arch. columns specify the architectural configurations for prediction and reconstruction modules, where MLP represents a single-hidden-layer MLP, 2MLP indicates two cascaded single-hidden-layer MLPs, and LB3M denotes our proposed architecture. Empty cells indicate the absence of corresponding tasks, while P.S. represents the parameter size normalized to a single MLP.

For reconstruction tasks, a single MLP structure (Line 2) exhibits limited capability with an I-AUROC of 0.919, showing notable degradation compared to the basic prediction task (0.954, Line 1). Our experiments reveal that utilizing two MLPs for reconstruction branch achieves comparable performance with an I-AUROC of 0.957 (Line 4), indicating that reconstruction tasks may require larger parameter capacity.

The basic Bidirectional Feature Mapping (BFM) framework with single-hidden-layer MLPs (CFM-Dual) achieves an I-AUROC of 0.962 with 4× parameters (Line 3). Building upon this enhanced reconstruction branch, using a single-hidden-layer MLP for prediction branch increases the parameter count to 6×, while using two single-hidden-layer MLPs for prediction further raises it to 8× parameters, as demonstrated in Lines 5 and 6. Although these configurations show moderate improvements over CFM-Dual (from 0.962 to 0.968 and 0.969), while benefiting from a better reconstruction branch, they introduce substantial parameter overhead.

Leveraging parameter sharing mechanism, LB3M achieves competitive performance with fewer parameters. While maintaining a modest parameter count (4× that of a single MLP), LB3M achieves superior performance with an I-AUROC of 0.974 and AUPRO@30% of 0.978 (Line 7). These results show that LB3M's architecture enables better feature pattern learning with fewer parameters than conventional approaches.

4.5. Few-shot anomaly detection

In many practical scenarios, collecting sufficient data for normal samples is challenging. Therefore, a model that performs well even with very limited data has a considerable advantage. To evaluate the few-shot learning capabilities of our model, we conduct experiments under 5-shot, 10-shot, 50-shot, and full-data settings.

As shown in Table 7, when a random subset of the training data is selected and the metric is averaged over five trials, our method consistently achieves the best results across all data volumes compared to methods that report performance under few-shot settings. Notably, the performance advantage of our method is most significant when fewer samples are available, highlighting its robustness in data-scarce conditions.

A comparative analysis between M3DM and CFM reveals that methods relying solely on inter-modal complementary relations, such as CFM, struggle to perform effectively in few-shot scenarios. In contrast, M3DM, which emphasizes intra-modal characteristics, demonstrates stronger performance with very limited samples. However, as the sample size increases, the advantage of single-modality analysis diminishes, and the benefits of inter-modal complementary relations become increasingly apparent.

Our method inherits the strengths of both approaches by jointly modeling intra-modal features and inter-modal correspondences. This integrated design, combined with self-supervised learning, enables our model to achieve superior performance in few-shot settings. The results demonstrate that our method is not only highly effective in data-scarce conditions but also adapts seamlessly as more data becomes available, showcasing robust few-shot learning capabilities.

5. Conclusion

In this work, we propose CPIR, a novel approach for anomaly detection that introduces the Bidirectional Feature Mapping (BFM) paradigm and the Latent Bridged Modal Mapping Module (LB3M) architecture. The BFM paradigm enables simultaneous modeling of intra-modal and inter-modal representations, while the LB3M design effectively integrates cross-modal prediction and intra-modal reconstruction tasks into a dual-task learning framework. Together, these innovations allow CPIR to capture comprehensive feature representations, significantly enhancing anomaly detection performance.

Extensive experiments validate the effectiveness of our method, with CPIR achieving state-of-the-art results on the MVTec 3D-AD dataset, excelling in both image-level and pixel-level metrics. Moreover, CPIR demonstrates robust performance in few-shot learning scenarios and

maintains a favorable balance between accuracy and computational efficiency. These results establish CPIR as a powerful and practical approach for anomaly detection, laying a solid foundation for future research in feature mapping-based methods and their applications.

CRediT authorship contribution statement

Wen Shangguan: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Hongqiang Wu:** Writing – review & editing, Visualization, Validation, Supervision, Conceptualization. **Yanchang Niu:** Writing – review & editing, Validation. **Haonan Yin:** Writing – review & editing, Validation. **Jiawei Yu:** Writing – review & editing, Methodology. **Bokui Chen:** Supervision. **Biqing Huang:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

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 was supported by the Science and Technology Program of Qingdao Municipality under grant 25-1-1-gjgg-32-gx.

Data availability

Data will be made available on request.

References

[1] 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.
[2] 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, 2025, pp. 75–91.
[3] A. Costanzino, P.Z. Ramirez, G. Lisanti, L. Di Stefano, Multimodal industrial anomaly detection by crossmodal feature mapping, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 17234–17243.
[4] 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. 2967–2976.
[5] 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, in: Proceedings of Machine Learning Research, vol. 202, PMLR, 2023, pp. 6185–6194.
[6] P. Bergmann, X. Jin, D. Sattlegger, C. Steger, The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization, 2021, arXiv preprint arXiv:2112.09045.
[7] 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.
[8] J. Liu, G. Xie, J. Wang, S. Li, C. Wang, F. Zheng, Y. Jin, Deep industrial image anomaly detection: A survey, Mach. Intell. Res. 21 (1) (2024) 104–135.
[9] V. Zavrtanik, M. Kristan, D. Skočaj, Dsr–a dual subspace re-projection network for surface anomaly detection, in: European Conference on Computer Vision, Springer, 2022, pp. 539–554.
[10] D. Gong, L. Liu, V. Le, B. Saha, M.R. Mansour, S. Venkatesh, A.v.d. Hengel, Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1705–1714.
[11] Y. Liang, J. Zhang, S. Zhao, R. Wu, Y. Liu, S. Pan, Omni-frequency channel-selection representations for unsupervised anomaly detection, IEEE Trans. Image Process. (2023).
[12] J. Song, K. Kong, Y.-I. Park, S.-G. Kim, S.-J. Kang, AnoSeg: anomaly segmentation network using self-supervised learning, 2021, arXiv preprint arXiv:2110.03396.
[13] X. Yan, H. Zhang, X. Xu, X. Hu, P.-A. Heng, Learning semantic context from normal samples for unsupervised anomaly detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 3110–3118.
[14] Z. You, L. Cui, Y. Shen, K. Yang, X. Lu, Y. Zheng, X. Le, A unified model for multi-class anomaly detection, Adv. Neural Inf. Process. Syst. 35 (2022) 4571–4584.
[15] A. De Nardin, P. Mishra, G.L. Foresti, C. Piciarelli, Masked transformer for image anomaly localization, Int. J. Neural Syst. 32 (07) (2022) 2250030.
[16] J. Wyatt, A. Leach, S.M. Schmon, C.G. Willcocks, Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 650–656.
[17] A. Mousakhan, T. Brox, J. Tayyub, Anomaly detection with conditioned denoising diffusion models, 2023, arXiv preprint arXiv:2305.15956.
[18] 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.
[19] 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.
[20] K. Batzner, L. Heckler, R. König, Efficientad: Accurate visual anomaly detection at millisecond-level latencies, 2023, arXiv preprint arXiv:2303.14535.
[21] M. Rudolph, B. Wandt, B. Rosenhahn, Same same but differnet: Semi-supervised defect detection with normalizing flows, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1907–1916.
[22] 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.
[23] S. Zagoruyko, N. Komodakis, Wide residual networks, 2016, arXiv preprint arXiv:1605.07146.
[24] A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[25] R. Yan, F. Zhang, M. Huang, W. Liu, D. Hu, J. Li, Q. Liu, J. Jiang, Q. Guo, L. Zheng, CAINNFlow: Convolutional block attention modules and invertible neural networks flow for anomaly detection and localization tasks, 2022, arXiv preprint arXiv:2206.01992.
[26] O. Rippel, A. Chavan, C. Lei, D. Merhof, Transfer learning gaussian anomaly detection by fine-tuning representations, 2021, arXiv preprint arXiv:2108.04116.
[27] N. Cohen, Y. Hoshen, Sub-image anomaly detection with deep pyramid correspondences, 2020, arXiv preprint arXiv:2005.02357.
[28] 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.
[29] H. Li, J. Hu, B. Li, H. Chen, Y. Zheng, C. Shen, Target before shooting: Accurate anomaly detection and localization under one millisecond via cascade patch retrieval, IEEE Trans. Image Process. (2024).
[30] Y. Shi, J. Yang, Z. Qi, Unsupervised anomaly segmentation via deep feature reconstruction, Neurocomputing 424 (2021) 9–22.
[31] Z. You, K. Yang, W. Luo, L. Cui, Y. Zheng, X. Le, Adtr: Anomaly detection transformer with feature reconstruction, in: International Conference on Neural Information Processing, Springer, 2022, pp. 298–310.
[32] H. Yin, G. Jiao, Q. Wu, B.F. Karlsson, B. Huang, C.Y. Lin, LafitE: Latent diffusion model with feature editing for unsupervised multi-class anomaly detection, 2023, arXiv:2307.08059.
[33] J. Hyun, S. Kim, G. Jeon, S.H. Kim, K. Bae, B.J. Kang, ReConPatch: Contrastive patch representation learning for industrial anomaly detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 2052–2061.
[34] 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.
[35] Y. Cao, X. Xu, W. Shen, Complementary pseudo multimodal feature for point cloud anomaly detection, Pattern Recognit. 156 (2024) 110761.
[36] 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 (2024).
[37] 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.
[38] Z. Zhou, L. Wang, N. Fang, Z. Wang, L. Qiu, S. Zhang, R3D-AD: Reconstruction via diffusion for 3D anomaly detection, in: European Conference on Computer Vision, Springer, 2024, pp. 91–107.
[39] H. Liang, G. Xie, C. Hou, B. Wang, C. Gao, J. Wang, Look inside for more: Internal spatial modality perception for 3D anomaly detection, 2024, arXiv preprint arXiv:2412.13461.
[40] H. Li, Y. Niu, H. Yin, Y. Mo, Y. Liu, B. Huang, R. Wu, J. Liu, DAUP: Enhancing point cloud homogeneity for 3D industrial anomaly detection via density-aware point cloud upsampling, Adv. Eng. Inform. 62 (2024) 102823.

[41] 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.
[42] R. Chen, G. Xie, J. Liu, J. Wang, Z. Luo, J. Wang, F. Zheng, EasyNet: An easy network for 3D industrial anomaly detection, 2023, arXiv preprint arXiv:2307.13925.
[43] V. Zavrtanik, M. Kristan, D. Skočaj, Draem-a discriminatively trained reconstruction embedding for surface anomaly detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8330–8339.
[44] A. Dai, A.X. Chang, M. Savva, M. Halber, T. Funkhouser, M. Nießner, Scannet: Richly-annotated 3d reconstructions of indoor scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5828–5839.
[45] 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.
[46] Y. Pang, W. Wang, F.E. Tay, W. Liu, Y. Tian, L. Yuan, Masked autoencoders for point cloud self-supervised learning, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II, Springer, 2022, pp. 604–621.
[47] 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.
[48] B. Ma, Z. Han, Y.-S. Liu, M. Zwicker, Neural-pull: Learning signed distance functions from point clouds by learning to pull space onto surfaces, 2020, arXiv preprint arXiv:2011.13495.
[49] B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, R.C. Williamson, Estimating the support of a high-dimensional distribution, Neural Comput. 13 (7) (2001) 1443–1471.
[50] J. Wang, X. Wang, R. Hao, H. Yin, B. Huang, X. Xu, J. Liu, Incremental template neighborhood matching for 3D anomaly detection, Neurocomputing 581 (2024) 127483.
[51] G. Qian, Y. Li, H. Peng, J. Mai, H. Hammoud, M. Elhoseiny, B. Ghanem, Pointnext: Revisiting pointnet++ with improved training and scaling strategies, Adv. Neural Inf. Process. Syst. 35 (2022) 23192–23204.
[52] 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.
[53] 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.
[54] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.
[55] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660.