Mentor3AD: Feature Reconstruction-based 3D Anomaly Detection via Multi-modality Mentor Learning

Jinbao Wang[1], Hanzhe Liang[1], Can Gao[1], Chenxi Hu[1], Jie Zhou[1], Yunkang Cao[2], Linlin Shen[1], Weiming Shen[3,†]
[1]Shenzhen University
[2]Hunan University
[3]Huazhong University of Science and Technology

***Abstract*—Multimodal feature reconstruction is a promising approach for 3D anomaly detection, leveraging the complementary information from dual modalities. We further advance this paradigm by utilizing multi-modal mentor learning, which fuses intermediate features to further distinguish normal from feature differences. To address these challenges, we propose a novel method called Mentor3AD, which utilizes multi-modal mentor learning. By leveraging the shared features of different modalities, Mentor3AD can extract more effective features and guide feature reconstruction, ultimately improving detection performance. Specifically, Mentor3AD includes a Mentor of Fusion Module (MFM) that merges features extracted from RGB and 3D modalities to create a mentor feature. Additionally, we have designed a Mentor of Guidance Module (MGM) to facilitate cross-modal reconstruction, supported by the mentor feature. Lastly, we introduce a Voting Module (VM) to more accurately generate the final anomaly score. Extensive comparative and ablation studies on MVTec 3D-AD and Eyecandies have verified the effectiveness of the proposed method.**

***Index Terms*—3D Anomaly Detection, point cloud, multimodal, contrastive learning.**

I. Introduction

THree Dimension Anomaly Detection (3DAD) has been extensively used in high-precision industrial product inspection, attracting considerable attention from the computer vision community [1]–[3]. It is dedicated to identifying anomalous points or regions that deviate from the normal distribution in a given 3D point cloud and depth data. Existing methods are mainly classified into unimodal 3DAD and multimodal 3DAD [4]–[7].

Unimodal 3DAD detects anomalies from the point cloud (or depth) structure, with methods based on feature-embedding [8]–[11] and point cloud reconstruction [1], [12]. However, these methods mainly focus on improving unimodal features and do not fully explore the complementarity of different modalities. Multimodal 3DAD enhances the feature set by integrating RGB and 3D modalities, including depth maps and point clouds. While depth maps can experience information loss due to occlusions, point clouds offer a more accurate representation of 3D structures. Effective fusion is critical for multimodal anomaly detection. However, existing methods suffer from interference between different modalities because of the insufficient fusion. Several methods have been proposed

Corresponding Author†. This paper is mainly realised by Hanzhe Liang at Shenzhen University, if you have any questions please contact: 2023362051@email.szu.edu.cn.

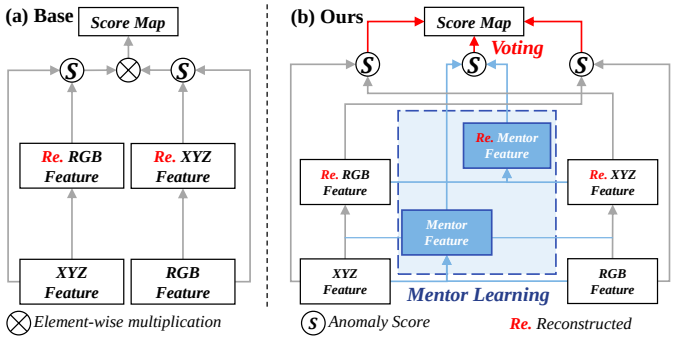


Fig. 1. Illustration of (a) the base mode and (b) our proposed mode. The base mode integrates features simply, while our mode excels in capturing shared features via mentor learning.

to fully utilize the complementary nature of these modalities. For example, BTF [13] and M3DM [14] provide basic fusion strategies; however, they neglect the combination mechanisms between different modalities, which leads to poor performance in the final scoring stage. Additionally, shape-guided [15] relies on feature alignment. Although this method attempts to leverage complementary information between modalities, it may not fully exploit it, which negatively impacts subsequent detection performance. Therefore, a key issue is posed: *how to effectively integrate multimodal features, that is, to enhance the discriminative fusion features and suppress the negative effects (e.g., false positives)?*

To address this question, based on the fact that the shared features in multiple modalities play a crucial role in enhancing the models' discriminative abilities, we present a novel approach called Multi-modality Mentor Learning (Mentor3AD) for detecting anomalies in multimodal data, as illustrated in Figure 1. Specifically, Mentor3AD targets two main challenges related to multimodal information fusion in multimodal mentor learning. **(1) Weak Representation.** Previous feature reconstruction models directly and reconstruct features from one modality to another. However, it is insufficient for handling complex feature maps. Besides, these models do not leverage the correlation information between different modalities, leading to suboptimal outcomes. At the same time, the difference between normal and abnormal feature maps is often not obvious, which negatively impacts the anomaly detection performance. **(2) Weak Discrimination.** When multiple modalities

are introduced, the feature reconstruction model struggles with discriminatory information from these modalities.

The proposed Mentor3ad consists of three modules to enhance feature fusion and improve the model's discrimination ability. The first is the Mentor of Fusion Module (MFM), which combines RGB and 3D features into the mentor features. The second is the Mentor of Guidance Module (MGM), which performs the cross-modality reconstruction facilitated by the shared mentor features. Lastly, the Voting Module (VM) aggregates the anomaly detection results from different modalities to produce a final anomaly score.

The main contributions are summarized as follows:

- This paper proposes a new multi-modality mentor learning framework, called Mentor3AD for the 3DAD task, significantly improving performance and suppressing negative effects.
- To make better use of modal information, we designed MFM for generating mentor features to guide feature reconstruction, and the MGM guided by MFM to generate opposing modes. A Voting Module combines results from different modalities to generate final anomaly scores.
- The Mentor3AD achieves significant results in both comparative experiments and ablation studies, showcasing the effectiveness of the proposed method.

II. Related Work

A. RGB Anomaly Detection

2D image anomaly detection comprises feature extraction and feature modeling [5], [16], [17]. Feature extraction aims to derive discriminative representations, while feature modeling captures the distribution of normal features to detect anomalies [18], [19]. Early methods employed autoencoders and inpainting frameworks [20]–[22]. Subsequent advancements integrated normalizing flows [23], [24] and memory banks [25] for robust density estimation. These innovations enhance 2D detection accuracy and extend to multimodal frameworks, advancing industrial inspection research.

B. Unimodal 3D Anomaly Detection

In unimodal 3DAD, several innovative methods address the challenges of defect identification in 3D point clouds, mainly classified into feature embedding and feature reconstruction methods. The feature embedding method determines anomalies by forming a normal feature distribution from the features of the training set, and by comparing the difference between the features to be tested and the normal feature distribution at the time of testing [25], [26]. Reg3D-AD [8] utilizes a registration-based approach and feature memory banks to preserve critical details essential for anomaly detection, though there may be challenges in feature extraction and registration dependency that could impact its robustness. To further constrain group-level features in Reg3D-AD, Group3AD [9] refines anomaly localization by employing group-level feature contrastive learning, which differentiates normal and abnormal patterns more effectively. Then ISMP [10] skillfully uses the internal view to align global and local features to fully mine the structural information. Looking3D [27] aligns 2D and 3D data for anomaly detection, particularly benefiting manufacturing and quality control tasks. The feature reconstruction method uses normal samples to train the model in its ability to reconstruct normal features, and identifies anomalies by the reconstruction error during testing. IMR-Net [1] eliminates potential anomalies by iteratively masking and reconstructing, and identifies anomalies by comparing reconstruction differences. R3D-AD [12] uses diffusion models to further improve the reconstruction accuracy of the model for better detection. PO3AD [6] obtains higher resolution anomaly detection by predicting point-level offsets. Moreover, real-time pose-agnostic methods such as SplatPose [28] and SplatPose++ [29], ensure efficient anomaly detection critical for industrial applications. Some zero-shot methods using LLM also achieve good results [30]–[33].

These methods obtain excellent detection performance on unimodal modes but face challenges when considering inter-modal complementarity. And utilising the complementarity of RGB and 3D point clouds might lead to more comprehensive anomaly detection.

C. Multimodal 3D Anomaly Detection

The landscape of multimodal 3DAD has been enriched by a variety of methods that aim to integrate different types of data for enhanced detection capabilities, which can be broadly categorized into feature embedding methods and feature reconstruction approaches. Feature embedding methods are represented by BTF, which highlights the importance of leveraging classical 3D features to identify defects, advocating for a focus on the foundational geometric properties of the data [13], AST [34], which employs an asymmetric student-teacher network architecture, where a normalizing flow teacher and a feed-forward student network collaborate to distinguish anomalies by creating a divergence in their outputs, and M3DM [14], which stands out with its hybrid feature fusion approach, demonstrating the benefits of combining multiple data modalities by utilizing RGB, XYZ, and fused features to create three memory banks for anomaly detection. Feature reconstruction methods are represented by Shape-Guided [15], which utilizes a dual-memory framework informed by shape information, making it particularly effective at identifying anomalies in both color and shape. Instead of employing fused modalities, the system utilizes shape features to guide the steering process, which may potentially result in a lack of more informative fused features when confronted with complex scenes. Another method, CFM [30], was proposed to align features across different modalities to improve the detection of abnormalities.

These methods collectively contribute to a more nuanced and effective approach to anomaly detection in 3D data. However, feature reconstruction methods in a multimodal context remain challenging, as evidenced by the difficulty of cross-modality reconstruction due to significant differences in feature distribution between modalities, leading to poor discrimination. This paper proposes an approach that uses mentor modality to address this problem, leading to better anomaly detection.

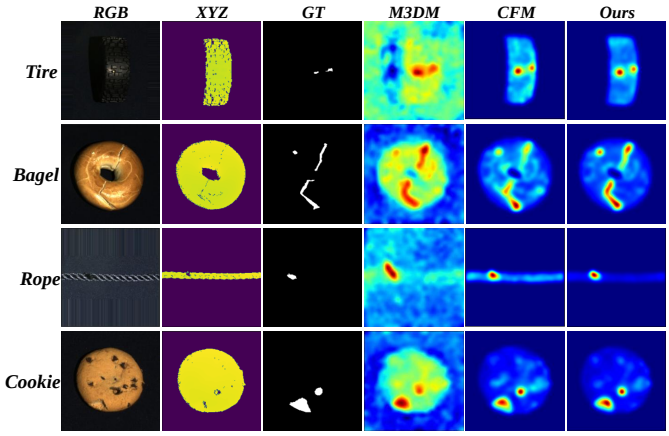


Fig. 2. More visualization results on MvTec 3D-AD. The distinction between anomalous and normal regions is more effective than previous methods. For instance, the normal region score of our method in the Rope class is nearly equivalent to zero, demonstrating its excellent anomaly detection performance.

III. Approach

A. Problem Statement

Multimodal anomaly detection (2D RGB + 3D point cloud) involves a training set defined as:

$$D^e_{\text{train}} = \left\{ \left(I_q \in \mathbb{R}^{H\times W\times 3},\ P_q \in \mathbb{R}^{N_q\times 3} \right) \right\}_{q=1}^{M},$$

which contains M normal samples from category e. Each sample consists of a 2D RGB image I_q (resolution $H \times W$, e.g., $H = W = 224$ in MvTec3D-AD and Eyecandies) and a 3D point cloud P_q (with N_q points). The test set is defined as: $D^e_{\text{test}} = \left\{ \left(I_q \in \mathbb{R}^{H\times W\times 3},\ P_q \in \mathbb{R}^{N_q\times 3},\ t_q \in T \right) \right\}_{q=1}^{K}$, where labels $t_q \in T = \{0, 1\}$ (0 for normal, 1 for anomaly). The objective is to train a deep anomaly detection model to build a scoring function: $\phi : \mathbb{R}^{H\times W\times 3} \times \mathbb{R}^{N_q\times 3} \to \mathbb{R}^{H\times W}$, for quantitatively evaluating the abnormality levels of new instances (combining RGB images and point clouds). We show several samples as shown in Figure 2.

B. Overview

The Mentor3AD method, illustrated in Figure 3, improves multimodal anomaly detection by employing a feature-based reconstruction that incorporates additional inter-modal mentor features. This approach allows our model to better use feature information and decision-making insights across different modalities. Consequently, the model becomes more robust and accurate in detecting anomalies in complex scenes.

Above all, the point cloud XYZ and RGB images are extracted into feature maps F_{XYZ} and F_{RGB} by the respective extractors. The proposed method is outlined as follows:

Training Phase. Contrastive learning is used to merge the shared features from the RGB and XYZ modalities into a low-dimensional mentor features, denoted as F_{Mtr}, through the Mentor of Fusion Module (MFM). After training the MFM, the parameter weights are frozen. The model then undergoes training using the Mentor of Guidance Module (MGM), which uses the capabilities of the mentor features to assist in cross-modal feature reconstruction. This involves 1) reconstructing F_{RGB} and F_{Mtr} to obtain $\tilde{F}_{XYZ}$, 2) F_{XYZ} and F_{Mtr} to obtain $\tilde{F}_{RGB}$, and 3) $\tilde{F}_{XYZ}$ and $\tilde{F}_{RGB}$ to obtain $\tilde{F}_{Mtr}$. Then we use the reconstructed differences of these three feature maps to train the Voting Module.

Test Phase. All weights are frozen, and both the F_{XYZ} and F_{RGB} feature maps are first input into the MFM to generate the F_{Mtr} feature maps. Then, the XYZ feature maps and the Mentor feature maps are fed into the pre-trained MGM to generate the reconstructed $\tilde{F}_{RGB}$ feature maps. Note that The reconstruction process for the RGB feature maps is the same as that of the XYZ feature maps. After this, the reconstructed $\tilde{F}_{RGB}$ and $\tilde{F}_{XYZ}$ feature maps are sent back into the MFM, which generates the reconstructed Mentor feature maps $\tilde{F}_{Mtr}$. By examine the differences in reconstruction among the three feature maps, three scoring maps are created. Finally, These scoring maps are then fed into the Voting Module (VM) to generate the anomaly scoring map.

C. Mentor of Fusion Module

It is essential to leverage the shared features between two modalities, especially when reconstructing a modal feature into another modality. Utilizing the shared information effectively can assist the model in gathering more features for detecting anomalies. Therefore, we propose a Mentor of Fusion Module (MFM), which uses an MLP framework to compress the dimensionality of the two modal feature maps. This process reduces the dimensionality of the fused modes to align with the model's requirements. The generated fused feature maps serve as guiding information, since they contain essential shared details common to both modalities, such as contours and shapes. Using the fused feature map as a mentor can help the model reconstruct the normal feature map more accurately while struggling with the abnormal feature map. This is because the fused model is effective at combining normal multimodal features but faces difficulties when dealing with abnormal modal features. This process enables a clearer distinction between normal and abnormal features. The inputs for the self-supervised learning of the mentor feature F_{Mtr} are the RGB feature map F_{RGB} and the point cloud feature map F_{XYZ}. The fusion process $MFM(F_{RGB}, F_{XYZ}) \to F_{Mtr}$ can be represented as follows:

$$F_{Mtr} = MLP\left(MLP\left(F_{RGB}\right) \oplus MLP\left(F_{XYZ}\right)\right). \quad (1)$$

The process involves aligning incoming bimodal features. This is accomplished by downscaling different dimensional feature maps from various modalities to a uniform dimension, followed by fusing these maps into a consolidated feature using a Multi-Layer Perceptron (MLP). The model ultimately receives aligned features through the function $MFM(F_{RGB}, F_{XYZ}) \to F_{Mtr}$. The next step in this process is to enhance the accuracy and detail of the information embedded in the aligned features by applying contrast loss. To self-supervise the learning of shared information between different modal feature maps, we use InfoNCE Loss for contrastive learning [14], [35]. This loss function encourages

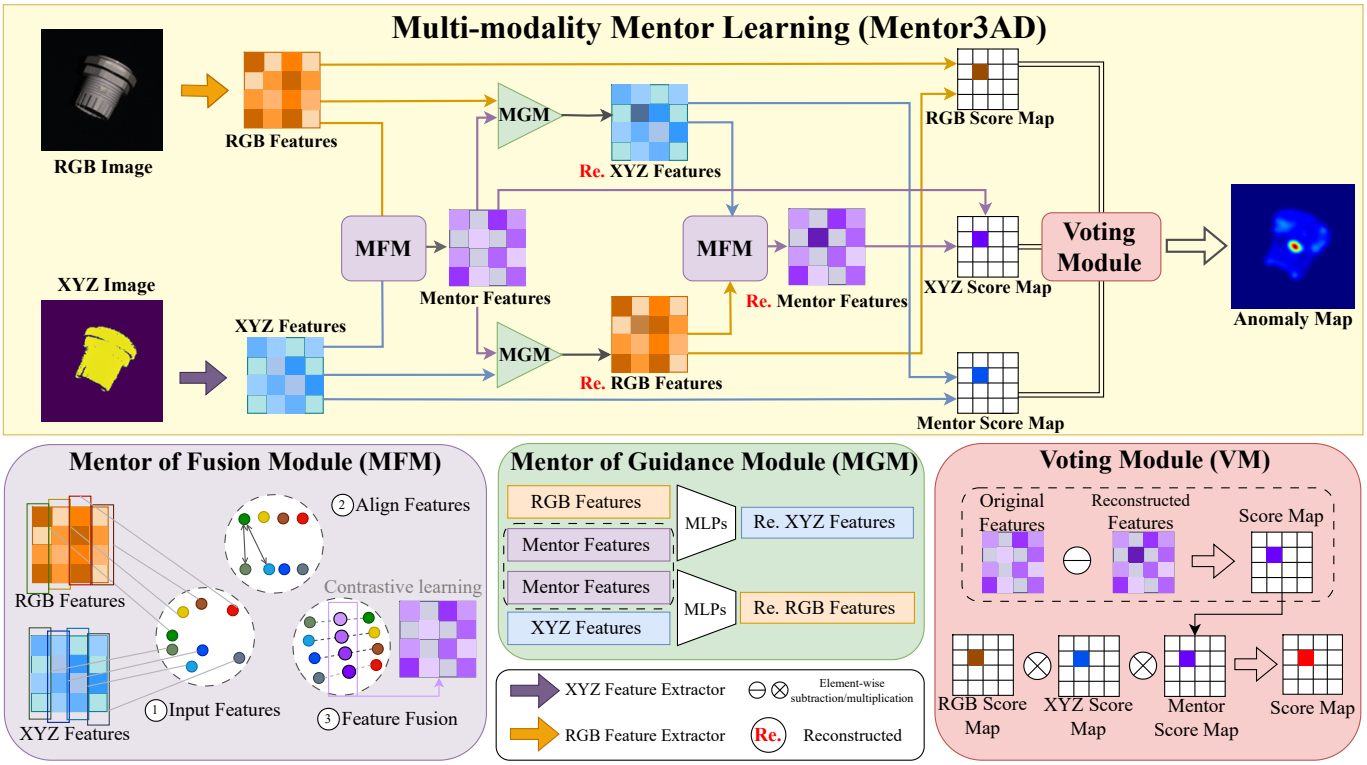


Fig. 3. The pipeline of Mentor3AD. **Training Phase:** MFM merges RGB and XYZ features into mentor features F_{Mtr}. MGM is used to reconstruct F_{RGB} and F_{Mtr}, yielding $\tilde{F}_{XYZ}$, while training another MGM, a process guided by the Mentor modal **Test Phase:** All weights are frozen. F_{XYZ} and F_{RGB} are input into MFM to generate F_{Mtr}, which are then used by MGM to reconstruct $\tilde{F}_{RGB}$ and $\tilde{F}_{XYZ}$. These are re-input into MFM to obtain $\tilde{F}_{Mtr}$. Reconstruction differences compute scoring maps, which are processed by the voting module (VM) to produce the final score.

the fusion of RGB modal feature maps with point cloud feature maps, facilitating a self-supervised approach to feature fusion. The loss function can be expressed as follows:

$$\mathcal{L}_{\text{con}} = \frac{F_{RGB}^{(i,j)} \cdot F_{XYZ}^{(i,j)}}{\sum_{t=1}^{N_b} \sum_{k=1}^{N_p} F_{RGB}^{(t,k)} \cdot F_{XYZ}^{(t,k)}}, \tag{2}$$

where N_b is the batch size and N_p is the nonzero patch number. In this context, let i represent the index of the training sample and j represent the index of the patch. By optimizing the loss function, we obtain the fused modal feature F_{Fusion}. This feature is subsequently used in the filter reconstruction module to guide the unimodal modes during the reconstruction process. The loss function was optimized with features shared between the RGB and XYZ modal feature maps, which were self-supervised and extracted in the form of a moderately dimensional fused feature map. This fused feature map plays an essential role in guiding the subsequent feature reconstruction.

D. Mentor of Guidance Module

In the study of multimodal feature map reconstruction, a typical method uses the normal feature map of one modality to reconstruct the normal feature map of another modality [30]. The core idea is to train the model to reconstruct using only normal feature maps, enabling the use of reconstruction error for anomaly detection during inference. However, challenges exist in accurately reconstructing normal features and effectively discriminating anomalies when directly mapping between modalities.

Therefore, we propose the Mentor of Guidance Module (MGM), which introduces a mentor modality into the model, guided by feature fusion, to address these challenges. When the guiding features are normal, the reconstructed feature maps are more accurate. Conversely, when the guiding features are abnormal, the reconstructed feature maps also display increased abnormalities. This occurs because the mentor modality is less effective at fusing abnormal feature maps; therefore, when it attempts to guide the reconstruction of these abnormal maps, it amplifies the reconstruction error further. Due to this mechanism, the introduction of the mentor modality not only enhances the model's reconstruction accuracy for normal features but also improves its ability to differentiate between normal and abnormal features.

We take the RGB modal F_{RGB} with the fused mentor modality F_{Mtr} as an example to reconstruct F_{XYZ}, and its result denotes $\tilde{F}_{XYZ}$. The reconstruction of F_{RGB} results in $\tilde{F}_{RGB}$ following the same logic. The process $MGM(F_{Mtr}, F_{RGB}) \rightarrow \tilde{F}_{XYZ}$ can be illustrated as follows: We use the RGB modality F_{RGB} along with the fused mentor modality F_{Mtr} as an example to reconstruct F_{XYZ}. The result of this reconstruction is denoted as $\tilde{F}_{XYZ}$. The reconstruction of F_{RGB} results in $\tilde{F}_{RGB}$ following the same logic. The process of $MGM(F_{Mtr}, F_{RGB}) \rightarrow \tilde{F}_{XYZ}$ can be illustrated

as follows:

$$\tilde{F}_{XYZ} = MLP\left(MLP\left(F_{Mtr}\right) \oplus F_{RGB}\right). \tag{3}$$

The processing of the mentor modality F_{Mtr} is performed using a single MLP. The features obtained from this processing are then combined with F_{RGB}. The combined features are further processed using three separate MLPs. The loss function $\mathcal{L}_{cos}$ is calculated based on the cosine similarity between the original feature F_{RGB} and the reconstructed feature $\tilde{F}_{RGB}$, as illustrated below:

$$\mathcal{L}_{cos} = 1 - \frac{\sum_{i=1}^{n} \tilde{F}_{XYZ,i} F_{XYZ,i}}{\sqrt{\sum_{i=1}^{n} \tilde{F}_{XYZ,i}^2} \sqrt{\sum_{i=1}^{n} F_{XYZ,i}^2}}, \tag{4}$$

where $\tilde{F}_{RGB,i}$ and $F_{RGB,i}$ represent the ith component of the vectors $\tilde{F}_{RGB}$ and F_{RGB}, respectively. n denotes the dimensionality of the feature vector. The loss function evaluates the level of dissimilarity between the two feature vectors.

This MGM enhances the ability of the model to differentiate between abnormal and normal states by establishing a triple distinction. First, the model operates by accepting a bi-modal feature map from the feature extractor to generate a mentor modality. If the pre-fusion feature map is determined to be abnormal, creating a mentor modality will further reinforce the distinction between abnormal and normal states. Second, an abnormal feature map is combined with an abnormal mentor modality to generate an additional abnormal modality, which helps to further differentiate between abnormal and normal states. Lastly, the RGB and XYZ eigenmaps reconstructed by the MGM are fed back into the MFM, which then predicts the mentor modality, denoted as $\tilde{F}_{Mtr}$. This step is akin to reconstructing the mentor modality. When the anomalous RGB and XYZ eigenmaps produced by the MGM are inputted, the already reconstructed anomalous eigenmaps are further enhanced, effectively widening the gap between normal and abnormal states.

E. Voting Module

Effectively leveraging the reconstruction differences among the three modalities is crucial. CFM [30] shows that multiplication can effectively capture the interactions between different scoring maps, largely due to the significant differences in their magnitudes. However, simple multiplicative methods may struggle to handle multiple anomaly scoring maps and might not provide optimal performance when integrating modalities with varying features. Furthermore, M3DM [36] demonstrates that directly using multiple One-Class Support Vector Machines to provide scores for each modality leads to the need to assign score weights to each One-Class Support Vector Machine in advance, which creates the challenge of finding optimal parameters.

To address this, we propose a Voting Module (VM) to understand better how different modalities—RGB, XYZ, and Mentor—contribute to anomaly detection. Similar to how we compute loss, we determine the anomaly score by calculating the cosine similarity between the original and reconstructed features. For instance, the RGB modal score S_{RGB}, can be computed as follows:

$$S_{RGB} = 1 - \frac{\sum_{i=1}^{n} \tilde{F}_{RGB,i} F_{RGB,i}}{\sqrt{\sum_{i=1}^{n} \tilde{F}_{RGB,i}^2} \sqrt{\sum_{i=1}^{n} F_{RGB,i}^2}}. \tag{5}$$

Using the same approach, we calculate the anomaly scoring maps S_{RGB}, S_{XYZ} and S_{Mtr}, and then S_{All} is calculated according to the following procedure:

$$S_{All} = \prod_{n=1}^{N} f\left({S_{RGB}}^{\alpha_n} \cdot {S_{XYZ}}^{\beta_n} \cdot {S_{Mtr}}^{\gamma_n}\right), \tag{6}$$

where $f(\cdot)$ aims to generate more refined scoring maps, providing more accurate scoring results for each pixel and enhancing the significance of anomalies in the ratings. Here, S_{RGB}, S_{XYZ}, and S_{Mtr} represent the anomaly scores in the RGB, XYZ, and mentor modalities, respectively. Besides, α_n, β_n, and γ_n are weighting exponents for different evaluation levels. These exponents are used to adjust the contribution of each modal disparity map to the final score. By concatenating and multiplying the results of these weighted disparity maps, we can derive a composite score S that reflects the overall evaluation across multiple reconstruction disparities. The function f can be expressed as follows:

$$f = C^U\left(C^L\left(S_{Input}\right)\right), \tag{7}$$

where S_{Input} represents the fraction to be calculated, C represents the convolution, and its superscripts U and L represent the convolution at different depths.

Then we use a learnable One-Class Support Vector Machine O_s to make the final anomaly segmentation map S'_{All}, which can be formalised as:

$$S'_{All} = O_s(S_{All}, \Theta) \tag{8}$$

where Θ stands for the parameters of O_s. The training process of O_s is shown in Algorithm 1. We train O_s through the score map S_{All} of the training set. Moreover, the final calculation result anomaly map S'_{All}, is used to calculate the score of each pixel. The object score is calculated by $Max(S'_{All})$ [25].

This voting module allows the model to make more effective use of the reconstruction differences between the three modes, thereby improving anomaly detection performance.

Algorithm 1 One-Class Support Vector Machine O_s Training

Output: Reconstructing Difference Map S_{All}, OCSVM layer O_s, OCSVM loss function L_{oc} [37].
Input: OCSVM parameters Θ.

1: **for** $s_{\text{all}} \in S_{\text{All}}$ **do**
2: $\quad \Theta \xleftarrow{\text{optim}} L_{oc}(O_s(s_{\text{all}}); \Theta)$ {Optimize parameters of O_s}
3: **end for**

IV. Experiment

Datasets. We conduct extensive experiments on MVTec 3D-AD [38] and Eyecandies [39]. MVTec 3D-AD is a multi-modal anomaly detection dataset containing both RGB and 3D structural information. It includes 4,147 sample pairs from 10 categories, of which 894 are anomalous. Eyecandies

	Method	Publication	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
I-AUROC	*Voxel Method (3D+RGB)*												
	VoxelGAN	*ICCV22*	0.680	0.324	0.565	0.399	0.497	0.482	0.566	0.579	0.601	0.482	0.517
	VoxelAE	*ICCV22*	0.510	0.540	0.384	0.693	0.446	0.632	0.550	0.494	0.721	0.413	0.538
	VoxelVM	*ICCV22*	0.553	0.772	0.484	0.701	0.751	0.578	0.480	0.466	0.689	0.611	0.609
	PointCloud Method (3D+RGB)												
	BTF	*CVPR23*	0.918	0.748	0.967	0.883	0.932	0.582	0.896	0.912	0.921	0.886	0.865
	AST	*WACV23*	0.983	0.873	0.976	0.971	0.932	0.885	0.974	0.981	**1.000**	0.797	0.937
	M3DM	*CVPR24*	0.994	**0.909**	0.972	0.976	0.960	**0.942**	0.973	0.899	0.972	0.850	0.945
	Shape-Guided	*CVPR24*	0.986	0.894	0.983	0.991	0.976	0.857	**0.990**	0.965	0.960	0.869	0.947
	CFM	*CVPR24*	0.994	0.888	0.984	0.993	0.980	0.888	0.941	0.943	0.980	**0.953**	0.954
	CFM-M	*CVPR24*	0.988	0.875	0.984	0.992	**0.997**	0.924	0.964	0.949	0.979	0.950	0.960
	Mentor3AD (XYZ+RGB)		0.992	0.900	0.982	0.994	0.995	0.900	0.980	**0.984**	**1.000**	0.910	0.964
	Mentor3AD		**0.996**	0.897	**0.988**	**0.995**	0.996	0.934	0.985	0.977	**1.000**	0.943	**0.971**
AUPRO@30%	*Voxel Method (3D+RGB)*												
	VoxelGAN	*ICCV22*	0.664	0.620	0.766	0.740	0.783	0.332	0.582	0.790	0.633	0.483	0.639
	VoxelA	*ICCV22*	0.467	0.750	0.808	0.550	0.765	0.473	0.721	0.918	0.019	0.170	0.564
	VoxelVM	*ICCV22*	0.510	0.331	0.413	0.715	0.680	0.279	0.300	0.507	0.611	0.366	0.471
	PointCloud Method (3D+RGB)												
	BTF	*CVPR23*	0.976	0.969	0.979	**0.973**	0.933	0.888	0.975	0.981	0.950	0.971	0.959
	AST	*WACV23*	0.970	0.947	0.981	0.939	0.913	0.906	0.979	0.982	0.889	0.940	0.944
	M3DM	*CVPR24*	0.970	0.971	0.979	0.950	0.941	0.932	0.977	0.971	0.971	0.975	0.964
	Shape-Guided	*CVPR24*	**0.981**	0.973	**0.982**	0.971	0.962	**0.978**	0.981	**0.983**	0.974	0.975	0.976
	CFM	*CVPR24*	0.979	0.972	**0.982**	0.945	0.950	0.968	0.980	0.982	0.975	0.981	0.971
	CFM-M	*CVPR24*	0.980	0.966	**0.982**	0.947	0.959	0.967	0.982	**0.983**	0.976	**0.982**	0.972
	Mentor3AD (XYZ+RGB)		**0.981**	0.965	0.920	0.951	0.950	**0.978**	0.982	**0.983**	0.981	0.980	0.967
	Mentor3AD		**0.981**	**0.976**	**0.982**	0.958	**0.966**	0.975	**0.983**	**0.983**	**0.982**	**0.989**	**0.978**
AUPRO@1%	*PointCloud Method (3D+RGB)*												
	BTF	*CVPR23*	0.428	0.365	0.452	0.431	0.370	0.244	0.427	0.470	0.298	0.345	0.383
	AST	*WACV23*	0.388	0.322	0.470	0.411	0.328	0.275	0.474	0.487	0.360	0.474	0.398
	M3DM	*CVPR24*	0.414	0.395	0.447	0.318	**0.422**	0.335	0.444	0.351	0.416	0.398	0.394
	CFM	*CVPR24*	0.459	**0.431**	0.485	0.469	0.394	0.413	0.468	0.487	0.464	0.476	0.455
	CFM-M	*CVPR24*	**0.480**	0.398	**0.490**	0.467	0.413	0.408	0.481	0.494	0.468	**0.488**	0.459
	Mentor3AD (XYZ+RGB)		0.478	0.402	0.487	**0.474**	0.396	**0.467**	0.488	**0.495**	**0.486**	0.476	0.465
	Mentor3AD		0.479	0.420	0.485	**0.474**	0.411	0.464	**0.498**	0.494	0.484	0.475	**0.468**

TABLE I

MAIN RESULTS ON MVTEC 3D-AD. I-AUROC(↑) EVALUATES THE MODEL'S ABILITY TO DETECT ANOMALIES AT THE SAMPLE LEVEL. P-AUPRO@30%(↑) AND P-AUPRO@1%(↑) EVALUATE THE MODEL'S ABILITY TO DETECT ANOMALIES AT THE PIXEL LEVEL UNDER GENERAL AND STRINGENT CONDITIONS, RESPECTIVELY. THE BEST AND THE SECOND-BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINE, RESPECTIVELY.

TABLE II

RESULTS ON THE EYECANDIES DATASET USING ONLY 350 TRAINING DATA. OUR METHOD WORKS BETTER USING LESS DATA TO CAPTURE MORE COMPLEX TRAINING STRUCTURES. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

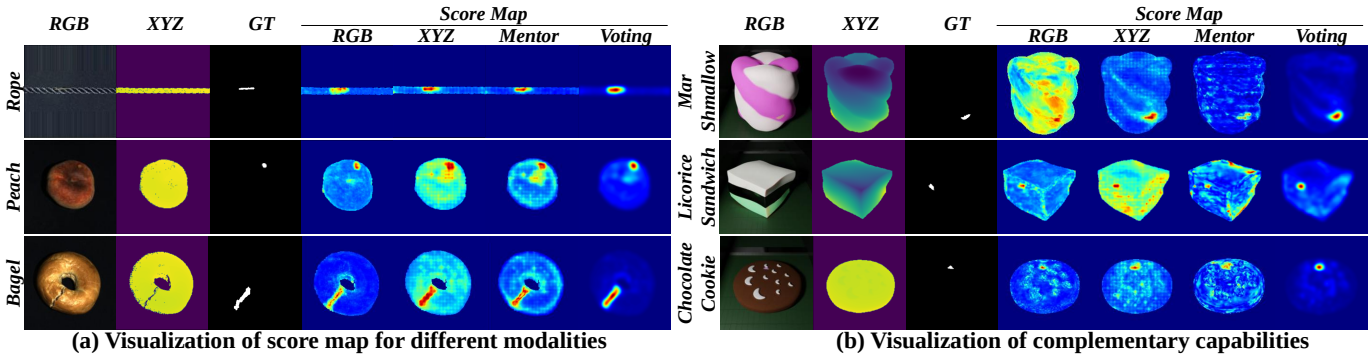
		Eyecandies										
	Method	Candy Cane	Chocolate Cookie	Chocolate Praline	Confetto	Gummy Bear	Hazelnut Truffle	Licorice Sandwich	Lollipop	Marshmallow	Peppermint Candy	Mean
I-AUROC	**BTF**	0.650	0.682	0.805	0.813	0.713	0.445	0.763	0.772	0.771	0.790	0.720
	M3DM	0.637	0.712	0.725	0.830	0.614	0.538	0.749	0.779	0.958	0.829	0.737
	CFM	0.661	**0.971**	**0.915**	0.939	**0.904**	**0.797**	0.850	0.879	**0.984**	0.877	0.878
	Ours	**0.688**	0.955	0.907	**0.952**	0.894	0.702	**0.925**	**0.893**	0.978	**0.896**	**0.879**
P-AUROC	**BTF**	**0.987**	0.914	0.917	0.921	0.838	0.817	0.884	0.957	0.897	0.811	0.894
	M3DM	0.975	0.962	0.926	0.989	0.889	0.835	0.955	0.943	0.993	0.982	0.945
	CFM	0.982	0.987	0.956	0.988	0.964	0.940	0.964	**0.977**	**0.995**	0.980	0.973
	Ours	0.981	**0.988**	**0.958**	**0.994**	**0.966**	**0.945**	**0.972**	**0.977**	0.993	**0.991**	**0.977**
P-AUPRO	**BTF**	0.938	0.739	0.700	0.707	0.656	0.470	0.663	0.882	0.719	0.619	0.709
	M3DM	0.925	0.825	0.725	0.956	0.659	0.456	**0.826**	0.704	**0.947**	0.910	0.793
	CFM	0.943	**0.894**	0.804	0.959	**0.855**	**0.781**	0.768	0.896	0.946	0.930	**0.878**
	Ours	**0.944**	0.843	**0.810**	**0.962**	0.840	0.779	0.799	**0.906**	0.939	**0.952**	0.877

TABLE III

THE FEW-SHOT RESULTS ON MVTEC 3D-AD. THE BEST AND THE SECOND-BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINE, RESPECTIVELY.

	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
Method	**I-AUROC**				**P-AUROC**				**AUPRO@30%**				**AUPRO@1%**			
BTF	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	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	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	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
Mentor3AD	**0.824**	**0.866**	**0.916**	**0.971**	**0.987**	**0.991**	**0.993**	**0.995**	**0.966**	**0.962**	**0.971**	**0.977**	0.345	**0.425**	**0.450**	**0.468**

contains 10 categories and 4,147 data pairs, 894 of which are anomalous. Eyecandies is also an RGB and 3D dataset containing 10,000 normal data pairs as training samples [39]. As existing methods use different benchmarks, e.g. some methods use only part of the normal data for training, while others use all of the data, this may have implications [14], [15], [30]. For a fair comparison, the number of training samples for each class is uniformly set to 349, giving a total of 3,500 training samples. Fewer samples show the method's excellent performance. The test samples are 250 normal data pairs and



(a) Visualization of score map for different modalities (b) Visualization of complementary capabilities

Fig. 4. Visualization analysis. (a) Visualization of score maps of each modality on MVTec 3D-AD. (b) Visualization of models with complementary capabilities on Eyecandies.

TABLE IV
RESULTS OF ABLATION EXPERIMENTS. w/o FOR NOT USING THIS MODULE, XYZ FOR XYZ SCORE MAP, RGB FOR RGB SCORE MAP AND Mtr FOR MENTOR SCORE MAP. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Method	I-AUROC	P-AUROC	P-AUPRO@30%	P-AUPRO@1%
$Mentor3AD_{\ w/o\ Vote\&Mtr}$	0.954	0.993	0.971	0.455
$Mentor3AD_{\ w/o\ XYZ\&Mtr}$	0.883	0.982	0.973	0.380
$Mentor3AD_{\ w/o\ RGB\&Mtr}$	0.906	0.982	0.977	0.397
$Mentor3AD_{\ w/o\ RGB\&XYZ}$	0.883	0.982	0.939	0.380
$Mentor3AD_{\ w/o\ Mtr}$	0.964	0.994	0.967	0.465
$Mentor3AD_{\ w/o\ Vote}$	0.967	0.990	0.964	0.450
Mentor3AD	**0.971**	**0.995**	**0.978**	**0.468**

250 anomalous data pairs. The experiment was divided into three parts. MVTec3D-AD was used for comparisons, with MVTec3D-AD for ablations, and samples from MVTec3D-AD and Eyecandies for few-shot experiments.

Methods. We select mainstream 3D+RGB and straightforward 3D methods, including VoxelGAN, VoxelAE, VoxelVM [38], BTF [13], AST [34], M3DM [14], Shape-Guided [15], CFM, and CFM-M [30]. All codes are derived from publicly available sources or published results, and their contributions are gratefully acknowledged.

Metrics. The Image Area Under the Receiver Operating Characteristic Curve (I-AUROC, ↑) is calculated using the global anomaly score to assess the performance of image-level anomaly detection. For pixel-level anomaly segmentation, the Area Under the Receiver Operating Characteristic Curve (P-AUROC, ↑) and the Area Under the Region Overlap (P-AUPRO, ↑) are utilized to evaluate performance. Additionally, AUPRO is examined at thresholds of 0.01 and 0.3, referred to as AUPRO@1% and AUPRO@30%, respectively, to analyze further the efficacy of pixel-level anomaly segmentation [30].

Implementation Details. In parallel with M3DM and CFM, the pre-trained weights of PointMAE are employed for 3D representation. In this process, a point cloud is transformed into 1,024 groups using FPS with KNN [14], [30], [40], [41]. Each group consists of 32 points and extracts features independently. The resulting features have a dimension of 1024×1152 and are ultimately projected onto an RGB image, creating a feature map that measures 224×224×1152 [42]. The frozen DINO VIT-B/8 model plays a crucial role in the characterization of the 224×224 image [43]. The image is carefully divided into 8×8 patches, allowing for the detailed characterization of each patch. This process produces a feature map with dimensions of 28×28×768. All training was conducted on a server equipped with a single NVIDIA A100-PCIE-40GB and a 64-core Intel Xeon Silver 4314 processor. To ensure consistent speed comparison criteria, tests were implemented on a server equipped with an RTX 4090 (24GB) and a Xeon(R) Platinum 8352V. All model performances are from publicly available papers. The CUDA version of one of the test processes was V11.3, the code was architected on Pytoch 1.10.0+cu113, and the Python version was 3.7.

A. Main Results

Comparison Results. We present the experimental results of our model on MVTec 3D-AD in Tables I. Our proposed method demonstrates a significant advantage in anomaly detection and segmentation compared to the previously leading 3D+RGB method. The I-AUROC has improved by 1.0%, reaching 97.1%, which reflects a notable enhancement in performance. The AUPRO@30% achieves performance comparable to the previous best method, with a SOTA performance of 97.8%. However, our method shows better inference efficiency, which will be discussed in subsequent sections. Additionally, the AUPRO@1% records 46.8% and performer better than the previous method under stringent criteria.

Efficiency Analysis. We evaluate our model's memory usage, inference speed, and overall performance, as outlined in Table V. By saving the final fusion results locally during pre-training, similar to the Shape-Guided approach, our model achieves improved efficiency. It also boasts a higher inference rate and lower memory usage compared to memory bank methods like M3DM, Shape-Guided, and BTF. Although it is slightly slower and uses more memory than CFM, our model surpasses it in 3DAD metrics, effectively balancing space-time efficiency with performance.

B. Ablation Studies

Analysis of Voting Module. We analyzed weighted combinations of modal score maps, as presented in Table VI.

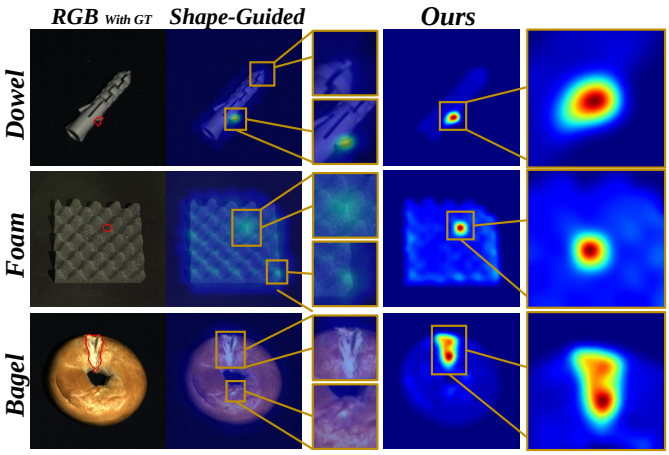


Fig. 5. Visualization of discriminative abilities. Previous methods, like Shape-Guided, often yield false positives, while our approach clearly distinguishes anomalies.

TABLE V
INFERENCE SPEED, MEMORY AND PERFORMANCE ON MVTEC 3D-AD. FRAME RATE IN FPS(↑) AND MEMORY IN MB(↓). HIGHER METRICS FOR 3DAD REPRESENT BETTER. THE BEST RESULTS IS HIGHLIGHTED IN **BOLD**.

Method	FrameRate	Memory	I-AUROC	P-AUROC	AUPRO@30%	AUPRO@1%
BTF	3.197	**381.1**	0.865	0.992	0.959	0.383
AST	4.966	463.9	0.937	0.976	0.944	0.398
M3DM	0.514	6526.1	0.945	0.992	0.964	0.394
Shape-Guided	1.513	1105.9	0.947	0.996	0.976	0.456
CFM	21.755	437.9	0.954	0.993	0.971	0.455
Mentor3AD	8.371	1615.3	**0.971**	**0.995**	**0.978**	**0.468**

TABLE VI
IMPACT OF DIFFERENT SCORING MANNERS. OUR VOTING PERFORMS BEST DUE TO THE VARYING SENSITIVITY OF DIFFERENT MODALITIES TO PIXEL-LEVEL VERSUS SAMPLE-LEVEL ANOMALIES.

Score Map	I-AUROC	P-AUROC	P-AUPRO@30%	P-AUPRO@1%
$f(XYZ \times RGB)$	0.965	0.994	0.973	0.464
$f(XYZ \times RGB \times Mtr)$	0.961	0.995	0.977	0.466
$f(XYZ \times RGB) \times f(Mtr)$	0.965	0.993	0.970	0.456
$f(Mtr \times RGB) \times f(XYZ)$	0.965	0.993	0.970	0.458
$f(Mtr \times XYZ) \times f(RGB)$	0.965	0.993	0.970	0.458
$f(XYZ) \times f(RGB) \times f(Mtr)$	0.967	0.990	0.964	0.450
Voting	**0.971**	**0.995**	**0.978**	**0.468**

Multiplying the three modality score maps improves pixel-level scoring, while multiplying individual score maps enhances sample-level scoring. Our experiments confirm that the proposed voting strategy, which balances scoring at both the sample and pixel levels, achieves the best results across all 3DAD metrics. This success is attributed to the complementary strengths of the different modalities: RGB detects surface anomalies, XYZ identifies structural anomalies, and the mentor modality Mtr combines both types of information.

Analysis of Each Module. We evaluated by removing each modality and voting strategy to assess validity, and the results are presented in Table IV. Our findings indicate that some subtle details may be overlooked when only RGB modality is used, and surface structure may not be effectively captured when only XYZ modality is available. Additionally, using fused mentor modality as complementary information to support 3DAD showed similar limitations. When these modalities are not utilized in combination, the results tend to be suboptimal. While using only XYZ and RGB modalities yields satisfactory results, they do not represent the best performance. A simple weighting of the XYZ, RGB, and mentor modalities can produce excellent detection results. Furthermore, incorporating the voting strategy leads to a significant improvement, achieving a 97.1% I-AUROC, which is considerably better than the previous best method using just XYZ and RGB. Our combined approach of XYZ, RGB, mentor modalities, and Voting achieves the best results.

C. More Experiments

Few-shot Results on MvTec3D-AD. The results for training set sizes of 5, 10, 50, and the full dataset, as shown in Table III, demonstrate that our method outperforms previous approaches across most metrics. Notably, it excels in pixel-level segmentation, even with limited samples. This fact can be attributed to the incorporation of fused modalities, which enhance the differentiation of anomalies compared to earlier methods that primarily focus on common features across modalities. The feature map-based reconstruction also outperforms memory bank-based methods because the latter struggles to represent the normal distribution with fewer samples, while the former is more efficient. Consequently, our method achieves superior performance in few-shot settings.

Few-shot Results on Eyecandies. Eyecandies is a challenging industrial synthetic dataset that provides 500 samples under various conditions. However, too many samples are difficult to obtain in the real industry, and there is a lack of uniform measurement criteria for existing models. We chose the first 350 samples as a smaller training set to test the performance of the model, and our performance achieves excellent results, with the SOTA P-AUROC and I-AUROC reaching 97.7% and 87.9%, respectively, as shown in Table II.

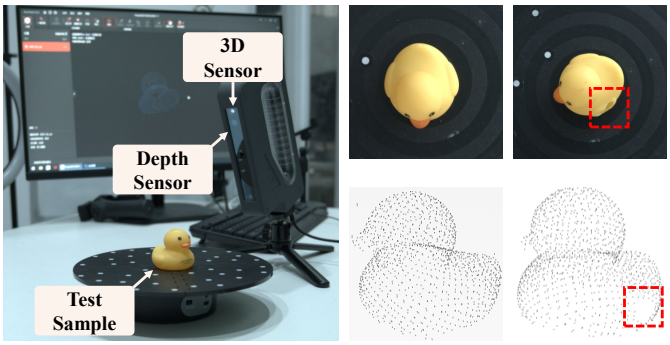
Feature Visualization. The results are shown in Figure 4. In part (a), we display the results of feature visualization for each modality of the MVTec 3D-AD sample, highlighting the differences in feature maps before and after reconstruction. Our model achieved excellent results in anomaly detection. Part (b) illustrates the results on Eyecandies, showcasing effective modal complementarity. Figure 5 shows that our model can better distinguish true and potential anomalies. Our model successfully localizes anomalies and assigns lower anomaly scores to normal pixels.

TABLE VII
QUANTITATIVE RESULTS FROM ACTUAL INDUSTRIAL PARTS DATASETS. RESULTS ARE EXPRESSED AS O-AUROC%/P-AUROC%. BEST RESULTS ARE IN BOLD.

Category	BTF	M3DM	Mentor3AD
Duck_1	71.0/72.3	83.3/76.2	**98.9/93.4**
Duck_2	76.2/60.4	68.3/59.9	**93.7/89.6**
Duck_3	63.7/74.3	71.0/83.4	**82.4/90.2**
Means	70.3/69.0	74.2/73.2	**91.7/91.1**

D. Actual Inspection on Industry Object

We conducted detection experiments on real industrial products to evaluate further the proposed model's actual performance in the real industry. The detection experiments include



(a) Scene (a) Data

Fig. 6. (a) The point cloud collection device of industry objects. (b) Some abnormal samples from four object classes.

three levels: 1) Duck_1: anomalies in 3D only, Duck_2: anomalies in RGB only, and Duck_3: 3D and RGB anomalies at the same time. Each class contains two training and 20 test sample pairs, each containing a point cloud and an RGB image. The 3D anomalies include bulges and concavities, and the RGB anomalies include painted colours. The scanning process is shown in Figure 6 (a), where the scanning process includes the object under test, the 3D scanning sensor and the depth scanning sensor. The data used are shown exemplarily in Figure 6 (b). To capture the corresponding RGB data, we used a NIKON Z6III with NIKKOR LENS Z 24-70mm f/4 S to shoot in the same pose. The quantitative results are presented in Table VII, where our model demonstrates excellent detection. Compared to the previous M3DM, our model improves 17.5% and 7.9% in anomaly detection and localization performance, respectively. Moreover, our model excels in RGB and 3D, which may be attributed to the mentor modality guiding the two modalities for cross-modal intermingling.

V. Conclusion

This paper presents a novel Multi-modality Mentor Learning (Mentor3AD) for detecting anomalies in multimodal 3DAD. Our method consists of three main modules that use the Mentor of Fusion Module (MFM) to combine RGB and 3D features into a single mentor modality, a Mentor of Guidance Module (MGM) that uses mentor modality to help reconstruct the modalities from each other, and a Voting Module (VM) that combines AD results from different modalities to generate a final anomaly score. Our model obtained the SOTA results, indicating the effectiveness of our method.

References

[1] W. Li, X. Xu, Y. Gu, B. Zheng, S. Gao, and 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 (CVPR)*, pp. 22207–22216, June 2024.

[2] Y. He, K. Song, Q. Meng, and Y. Yan, "An end-to-end steel surface defect detection approach via fusing multiple hierarchical features," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1493–1504, 2020.

[3] Z. Zhou, J. Wang, Z. Yu, Z. Wang, X. Liu, L. Qiu, and S. Zhang, "Featdae: Introducing features with denoising autoencoder for anomaly detection," *IEEE Transactions on Instrumentation and Measurement*, pp. 1–1, 2025.

[4] Y. Lin, Y. Chang, X. Tong, J. Yu, A. Liotta, G. Huang, W. Song, D. Zeng, Z. Wu, Y. Wang, and W. Zhang, "A survey on rgb, 3d, and multimodal approaches for unsupervised industrial image anomaly detection," *Information Fusion*, vol. 121, p. 103139, 2025.

[5] J. Liu, G. Xie, J. Wang, S. Li, C. Wang, F. Zheng, and Y. Jin, "Deep industrial image anomaly detection: A survey," *Machine Intelligence Research*, vol. 21, p. 104–135, Jan. 2024.

[6] J. Ye, W. Zhao, X. Yang, G. Cheng, and K. Huang, "Po3ad: Predicting point offsets toward better 3d point cloud anomaly detection," 2024.

[7] H. Liang, A. Wang, J. Zhou, X. Jin, C. Gao, and J. Wang, "Examining the source of defects from a mechanical perspective for 3d anomaly detection," 2025.

[8] J. Liu, G. Xie, X. Li, J. Wang, Y. Liu, C. Wang, F. Zheng, *et al.*, "Real3d-ad: A dataset of point cloud anomaly detection," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[9] H. Zhu, G. Xie, C. Hou, T. Dai, C. Gao, J. Wang, and L. Shen, "Towards high-resolution 3d anomaly detection via group-level feature contrastive learning," in *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, p. 4680–4689, ACM, Oct. 2024.

[10] H. Liang, G. Xie, C. Hou, B. Wang, C. Gao, and J. Wang, "Look inside for more: Internal spatial modality perception for 3d anomaly detection," 2025.

[11] Y.-Q. Cheng, W.-L. Li, C. Jiang, D.-F. Wang, H.-W. Xing, and W. Xu, "Mvgr: Mean-variance minimization global registration method for multiview point cloud in robot inspection," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–15, 2024.

[12] Z. Zhou, L. Wang, N. Fang, Z. Wang, L. Qiu, and S. Zhang, "R3d-ad: Reconstruction via diffusion for 3d anomaly detection," in *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXVI*, (Berlin, Heidelberg), p. 91–107, Springer-Verlag, 2024.

[13] E. Horwitz and 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*, pp. 2968–2977, 2023.

[14] Y. Wang, J. Peng, J. Zhang, R. Yi, Y. Wang, and C. Wang, "Multimodal industrial anomaly detection via hybrid fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8032–8041, 2023.

[15] Y.-M. Chu, C. Liu, T.-I. Hsieh, H.-T. Chen, and T.-L. Liu, "Shape-guided dual-memory learning for 3d anomaly detection," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 6185–6194, 2023.

[16] X. Xu, Y. Wang, Y. Huang, J. Liu, X. Lei, G. Xie, G. Jiang, and Z. Lu, "A survey on industrial anomalies synthesis," 2025.

[17] J. Wang, J. Cheng, C. Gao, J. Zhou, and L. Shen, "Enhanced fabric defect detection with feature contrast interference suppression," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–12, 2025.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[20] V. Zavrtanik, M. Kristan, and D. Skočaj, "Draem – a discriminatively trained reconstruction embedding for surface anomaly detection," 2021.

[21] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, June 2020.

[22] C. Gao, X. Chen, J. Zhou, J. Wang, and L. Shen, "Open-set fabric defect detection with defect generation and transfer," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–13, 2025.

[23] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Fully convolutional cross-scale-flows for image-based defect detection," in *Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2022.

[24] D. Gudovskiy, S. Ishizaka, and 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 (WACV)*, pp. 98–107, January 2022.

[25] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," 2021.
[26] Y. Cao, X. Xu, and W. Shen, "Complementary pseudo multimodal feature for point cloud anomaly detection," *Pattern Recognition*, vol. 156, p. 110761, 2024.
[27] A. Bhunia, C. Li, and H. Bilen, "Looking 3d: Anomaly detection with 2d-3d alignment," 2024.
[28] M. Kruse, M. Rudolph, D. Woiwode, and B. Rosenhahn, "Splatpose & detect: Pose-agnostic 3d anomaly detection," June 2024.
[29] Y. Liu, Y. S. Hu, Y. Chen, and J. Zelek, "Splatpose+: Real-time image-based pose-agnostic 3d anomaly detection," 2024.
[30] A. Costanzino, P. Zama Ramirez, G. Lisanti, and L. Di Stefano, "Multimodal industrial anomaly detection by crossmodal feature mapping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
[31] Q. Zhou, J. Yan, S. He, W. Meng, and J. Chen, "Pointad: Comprehending 3d anomalies from points and pixels for zero-shot 3d anomaly detection," in *Advances in Neural Information Processing Systems* (A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, eds.), vol. 37, pp. 84866–84896, Curran Associates, Inc., 2024.
[32] Y. Cheng, Y. Cao, G. Xie, Z. Lu, and W. Shen, "Towards zero-shot point cloud anomaly detection: A multi-view projection framework," 2024.
[33] Z. Zuo, J. Dong, Y. Wu, Y. Qu, and Z. Wu, "Clip3d-ad: Extending clip for 3d few-shot anomaly detection with multi-view images generation," 2024.
[34] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Asymmetric student-teacher networks for industrial anomaly detection," Jan. 2023.
[35] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019.
[36] C. Wang, H. Zhu, J. Peng, Y. Wang, R. Yi, Y. Wu, L. Ma, and J. Zhang, "M3dm-nr: Rgb-3d noisy-resistant industrial anomaly detection via multimodal denoising," 2024.
[37] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.
[38] P. Bergmann, X. Jin, D. Sattlegger, and 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*, SCITEPRESS - Science and Technology Publications, 2022.
[39] L. Bonfiglioli, M. Toschi, D. Silvestri, N. Fioraio, and D. De Gregorio, "The eyecandies dataset for unsupervised multimodal anomaly detection and localization," in *Proceedings of the 16th Asian Conference on Computer Vision (ACCV)*, 2022.
[40] Y. Pang, W. Wang, F. E. Tay, W. Liu, Y. Tian, and 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*, pp. 604–621, Springer, 2022.
[41] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," 2017.
[42] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16259–16268, 2021.
[43] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *CoRR*, vol. abs/2104.14294, 2021.