Highlights

MADFlow: Multimodal Difference Compensation Flow for Multimodal Anomaly Detection

Yao Li[#,],Xinyuan Zhou[#,],Shiyong Lan,Wenwu Wang,Xin Lai,Yixin Qiao

- We first introduce a normalizing flow-based framework for multimodal anomaly detection in point-cloud and image.
- We propose a **C**ross-modal **D**ifference **C**ompensation **F**usion module (**CDCF**) to align the features of the two modalities and compensate for the missing 3D information in the RGB image by using the difference between the aligned features.
- We propose a **F**requency-**S**pace **E**nhancement (**FSE**) module to improve the representation of the features from both frequency and space perspectives, thereby enhancing its ability to distinguish anomalies.
- We perform comprehensive experiments on both the MVTec 3D-AD and Eyecandies datasets, and show that the proposed method achieves the competitive performance.

MADFlow: Multimodal Difference Compensation Flow for Multimodal Anomaly Detection*

Yao Li[#,a], Xinyuan Zhou[#,a], Shiyong Lan[a,*], Wenwu Wang[b], Xin Lai[c] and Yixin Qiao[a]

[a]College of Computer Science, Sichuan University, Chengdu, 610065, China
[b]Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK
[c]School of Mechanical Engineering, Southwest Petroleum University, Chengdu, 610500, China

ARTICLE INFO

Keywords:
Anomaly detection
Normalizing flow
Feature fusion
Density estimation
Frequency-spatial feature enhancement

ABSTRACT

Normalizing flow-based methods have been widely studied in image anomaly detection and have proven their effectiveness. However, relying solely on 2D images is difficult to capture anomalies of objects from different perspectives. To address this issue, we propose **MADFlow**, where a **C**ross-modal **D**ifference **C**ompensation **F**usion (**CDCF**) module is designed to utilize 3D information and avoid excessive domain gap between multimodal features. Firstly, we perform consistent learning from features of different modalities by using a specific loss function. Based on this, the residual differences between the features of each modality are used to compensate for the missing 3D information in the 2D RGB data. In addition, we propose a **F**requency-**S**pace **E**nhancement (**FSE**) module, which models features from both frequency and space perspectives, and fuses them adaptively through a new gating mechanism, thus offering advantages over previous methods in modeling only spatial features. Finally, extensive experiments demonstrate the competitive results of the proposed method on two commonly used multimodal anomaly detection datasets, MVTec 3D-AD and Eyecandies.

1. Introduction

Since the release of the MVTec AD dataset [2], considerable progress has been made in industrial anomaly detection using 2D images [46; 34; 38; 37]. However, in real-world applications, RGB images are inherently sensitive to environmental factors, especially variations in illumination conditions, which can significantly hinder the reliable detection of anomalies. To address this challenge, multimodal anomaly detection (MAD) has emerged, combining RGB images with 3D structural data to enhance visual analysis [3]. The incorporation of 3D data provides geometric context, compensating for the shortcomings of RGB-only methods. As illustrated in Fig. 1, small and inconspicuous anomalies, such as those in the rope and peach classes, can be difficult to detect in RGB images alone but become more apparent when point clouds are integrated. The advancement of multimodal anomaly detection has been further supported by the release of the MVTec 3D-AD [3] and Eyecandies [5] datasets.

However, collecting a wide range of abnormal samples is highly challenging, as anomalies represent only a small fraction of real-world data. As a result, unsupervised methods have increasingly become the preferred approach for anomaly detection. According to [56], we can approximately categorize the current unsupervised multimodal methods into reconstruction-based [4; 7], memory-based [54; 56], distillation-based [14; 20], and normalizing flow-based anomaly detection [27; 45; 48] .

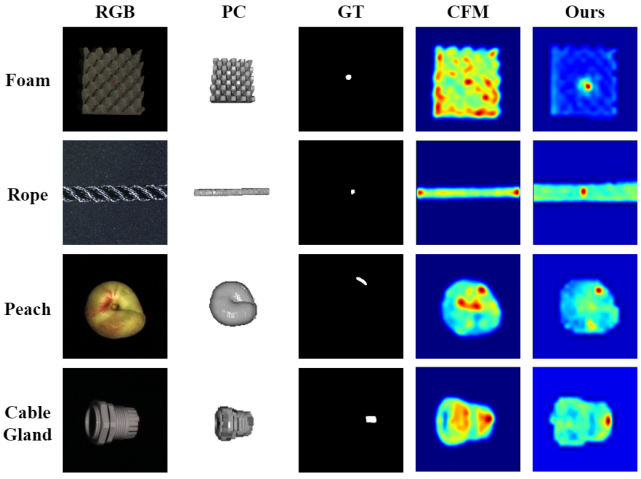

Figure 1: Illustration of the MVTec 3D-AD dataset [3]. The first column and the second column represent the input RGB image and point-cloud, respectively. The third column represents the ground truth, the fourth column is the anomaly detection heat map of the latest SOTA method CFM [10], and the fifth column is the heat map of our proposed method.

Reconstruction-based methods [4; 7] aim to reconstruct normal samples and use the reconstruction error to differentiate between normal and abnormal data. However, point clouds typically contain thousands of points with millimeter-level precision, making their accurate reconstruction highly complex. Since depth images can be considered a single-view projection of point clouds, many existing reconstruction methods opt to use depth images as a substitute for point clouds [7; 4]. However, in real-world scenarios, the location of anomalies is unpredictable. As a result, depth images may contain blind spots, which can negatively impact detection performance.

*This research is supported by National Natural Science Foundation of China project 62371324. This work is also supported by 2035 Innovation Pilot Program of Sichuan University, China. The codes are available at https://github.com/SYLan2019/MADFlow.
*Corresponding author: lanshiyong@scu.edu.cn. # Equal contribution.

Memory-based methods [54; 56] for multimodal anomaly detection typically use a pre-trained feature extractor to obtain features from normal samples, which are then stored in a memory bank. Sampling techniques [46] are used to reduce memory consumption while preserving the diversity of these features. However, this approach is limited by its strong reliance on the performance of the feature extractor.

Distillation-based anomaly detection methods [14; 20] build a teacher network and a student network, using the discrepancy between their outputs as the basis for anomaly detection. However, the existing method [20] uses depth images, and has not explored the use of point cloud data.

Normalizing flow-based anomaly detection methods [48; 45; 27] are often built on normalizing flow models, which can map any complex distribution of samples in a dataset to a Gaussian distribution by constructing a series of affine coupling layers [45; 27]. Such methods have been widely studied and proven to be effective in various fields, such as for anomaly detection from images [64; 27; 31], videos [8], and time series [23; 22]. However, currently there is no normalizing flow-based anomaly detection method that processes both images and point clouds.

To fill this gap, we explore the application of normalizing flow in multimodal anomaly detection, where each sample contains data from two modalities, namely, RGB image and point clouds. Since the data formats of these two modalities are completely different, directly concatenating them can lead to limited performance. To mitigate this issue, cross-modal feature mapping (CFM) [10] performs anomaly detection by learning cross-modal relationships between the features, and then uses the reconstruction error to detect anomalies. However, we argue that the differences between the modalities following the cross-modal alignment still carry valuable information for anomaly detection. These differences not only represent the features of individual samples but also provide complementary insights, which are particularly beneficial in scenarios where anomaly information is limited. Specifically, a Cross-Modal Alignment (CMA) module can be introduced to align the features of the images and point clouds. The resulting aligned differences, which capture unaligned cross-modal information (such as 3D geometric anomalies and local mismatches caused by anomalies), can be used as compensation information to enhance the feature representations of images.

Inspired by this, we propose a **C**ross-modal **D**ifference **C**ompensation **F**usion (**CDCF**) module, which mainly contains a **C**ross-**M**odal **A**lignment (**CMA**) component, to align the features from images and point clouds. The difference between these aligned features is used as a compensation feature to enrich the 3D structural information representation. The output of the CDCF module is then passed to a normalizing flow module, providing a more accurate fused representation of multimodal sample data.

In addition, given the diversity in the type of anomalies, it is challenging for normalizing flows to capture a wide range of feature representations. Although CNN-based spatial feature modeling is commonly used, it is constrained by a limited receptive field, making it less practical for capturing global spatial features. Fortunately, in terms of the convolution theorem, convolution in the spatial domain can be transformed into point multiplication in the frequency domain, allowing efficient modelling of global spatial relationships. In addition, the frequency domain processing offers complementary information that may be difficult to capture directly in the spatial domain, providing a valuable alternative perspective. Therefore, we propose a **F**requency-**S**pace **E**nhancement (**FSE**) module within each normalizing flow block, which enhances feature representations from both frequency and spatial perspectives.

In summary, our contributions are mainly as follows

- We first introduce a normalizing flow-based framework for multimodal anomaly detection (MADFlow) from point clouds and images.
- We propose a **C**ross-modal **D**ifference **C**ompensation **F**usion (**CDCF**) module to align the features of the two modalities and compensate for the missing 3D information in the RGB image by using the difference between the aligned features.
- We propose a **F**requency-**S**pace **E**nhancement (**FSE**) module to improve the representation of the features from both frequency and space perspectives, thereby enhancing its ability to distinguish anomalies.
- We perform comprehensive experiments on both the MVTec 3D-AD and Eyecandies datasets, and show that the proposed method achieves competitive performance.

2. Related Work

2.1. 2D Anomaly Detection

2D anomaly detection has been widely studied, benefiting from the continuous development of deep learning. We can categorize the methods for 2D anomaly detection as follows.

Reconstruction-based anomaly detection represents the most widely used method in the field of anomaly detection. Early approaches to 2D unsupervised anomaly detection primarily relied on generative models, such as variational autoencoders (VAEs) [28], generative adversarial networks (GANs) [19], and diffusion models [24]. These methods are trained exclusively on normal samples, and anomalies are detected during inference by measuring the difference between the input and its reconstruction. However, a common limitation of such models is their tendency to generalize, enabling them to partially reconstruct subtle anomalies even when trained only on normal data, thereby degrading anomaly detection performance. To address this issue, several methods [32; 35; 59] introduce artificially generated anomalies into the training set. Although this approach helps the model learn abnormal patterns, synthetic anomalies cannot capture the full diversity of real-world defects, leading to overfitting and reduced real-world detection performance.

Memory-based anomaly detection methods extract features from training samples, often using models pre-trained on large-scale datasets such as ImageNet [15], and store the features of normal samples in a memory bank. During inference, features from a test sample are compared with those in the memory bank, with the distance to the most similar feature used as the anomaly score. However, this approach requires storing a large number of normal sample features. To reduce memory usage while preserving feature diversity, the work in [46] proposes a greedy core subset sampling method. Since the anomaly score is obtained by computing the distance between the feature of interest and all sample features in the memory bank, the selection of stored samples is critical for both detection performance and computational efficiency.

Distillation-based anomaly detection assumes that the teacher network is capable of extracting both normal and abnormal features [61]. In contrast, the student network is trained exclusively on normal data, learning only to replicate the teacher's outputs for normal samples. As a result, during testing, the student exhibits a noticeable discrepancy from the teacher when processing anomalous inputs [61]. This difference serves as an effective indicator for detecting anomalies. However, since both networks typically follow the same data flow, they may still produce similar outputs. To address this, recent methods [14; 51] introduce a reverse distillation strategy, reversing the data flow to amplify differences and improve anomaly detection.

Normalizing flow-based anomaly detection has received widespread attention in recent years. These methods map the features of normal samples to a simple, tractable distribution, treating samples that deviate from this distribution during testing as anomalies [64; 27; 31]. However, as noted in [30], applying the normalizing flow directly to RGB images tends to assign a high probability value to abnormal images, leading to false detection of anomalies. In [47], this problem is alleviated by applying the normalizing flow to high-dimensional features. In 2D anomaly detection, normalizing flow is increasingly used to improve the separation between normal and abnormal samples [64; 31; 47]. A common strategy is to incorporate multi-scale information. For example, the method in [47] maps features at multiple scales into a unified distribution to boost detection across varying anomaly sizes. The work in [31] enhances the model's detection performance by enhancing the interaction between features at different scales. In [64], multiple flow models are used to handle multi-scale feature dependencies. However, these multi-scale designs can significantly increase computational cost, limiting real-time applicability.

2.2. 3D Anomaly Detection

Due to the inherent limitations of image-based anomaly detection, such as sensitivity to lighting conditions, there is a growing interest in multimodal anomaly detection. While multimodal fusion has been extensively explored in various tasks [1; 17; 18; 22; 36], most of these methods are designed for general classification or fusion purposes. In contrast, anomaly detection places greater emphasis on capturing and leveraging the discrepancies between different modalities rather than simply combining them.

Reconstruction-based methods aim to reconstruct both modalities to capture rich feature representations. To improve anomaly detection, approaches such as EasyNet [7] and Dual-branch Reconstruction [4] introduce synthetic anomalies into normal data during training. However, directly simulating anomalies in complex point-cloud data is difficult, so many methods instead use depth images as a substitute. EasyNet specifically addresses a key limitation of this substitution: when an anomaly appears in only one modality, it may be masked by accurate reconstruction in the other. To address this limitation, EasyNet introduces an information entropy-based fusion strategy that adaptively weights and combines anomaly scores from each modality. However, this approach follows a late fusion paradigm, preventing the model from fully learning the interactions between modalities.

Distillation-based methods focus on transferring normal patterns while suppressing anomalous ones. MMRD [20], for example, uses a frozen teacher network to generate targets, while the student network learns to reconstruct features of normal samples and detect anomalies through reconstruction errors, effectively capturing local deviations. However, this non-parametric fusion approach lacks a clearly defined learning objective. In addition, the student network in MMRD relies on retrieving and generating priors from a fixed set of prototypes, which can limit its generalization ability if the training data do not sufficiently represent the full range of normal conditions.

Memory-based methods build a memory bank for images and point clouds, respectively [56; 54]. During testing, the feature distances between each modality and all entries in the memory bank are computed independently. The anomaly score is then determined based on the difference between the test feature and its closest match in the memory bank. To maintain a balance between memory efficiency and feature diversity, the memory bank is constructed using greedy core-set sampling [46].

Normalizing flow-based methods have been rarely explored in 3D anomaly detection, primarily due to the challenge of effectively fusing features from images and point clouds. Although such methods [47; 64; 27] have been widely applied to 2D image anomaly detection, their use in point cloud processing has mostly focused on generation tasks [58; 42]. Unlike traditional approaches such as VAE [28] or GAN [19], which depend on fixed point sampling, thus limiting the quality of generation, normalizing flows can model the spatial distribution of diverse shapes and learn the surface point distribution more flexibly. However, since point clouds contain thousands of points, they allow for some imprecision in local details during generation. In contrast, anomaly detection is highly sensitive to such local variations, making it more challenging to apply normalizing flows effectively. Existing methods like [48] replace point clouds with depth images, but this loses structural information and leads to suboptimal performance. Our approach seeks

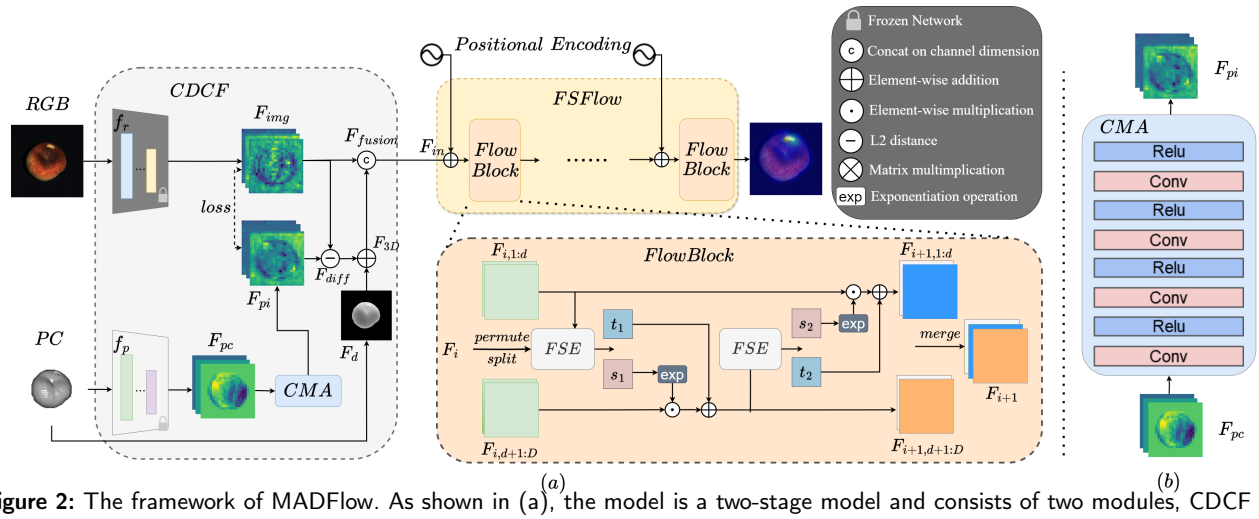

Figure 2: The framework of MADFlow. As shown in (a), the model is a two-stage model and consists of two modules, CDCF and FSFlow. In stage 1, we first train the CMA network to align image features with point-cloud features, and then fuse the difference features with the depth features to obtain the fused features. In stage 2, FSFlow, which is composed of multiple flow blocks, is used to map the fused features of the CDCF to a Gaussian distribution. In each flow block, we propose FSE to enhance the features from both frequency and spatial perspectives. The detail of **C**ross-**M**odal **A**lignment (**CMA**) is shown in (b), which is composed of four convolutions and activation functions.

to bridge this gap by leveraging full point-cloud data for improved anomaly detection.

3. Proposed Method

3.1. Problem Formulation and Background

We first give the definition of the problem. Given a set of anomaly-free training samples $T = \{(\mathrm{I}_i, \mathrm{P}_i)\}_{i=1}^{N_s}$, where I_i and P_i represent the i-th RGB image and point cloud, respectively, and N_s represents the number of training samples, our goal is to train an anomaly detection model that can accurately distinguish anomalies in a test set containing normal and abnormal samples.

Before presenting our method, we briefly introduce the normalization flow, which forms the basis of our method. Unlike other generative models, such as the VAE [28] and GAN [19], normalizing flow has the unique ability to progressively transform any sample distribution, including multi-peak ones, into a standard Gaussian distribution. This is achieved by stacking a sequence of affine coupling layers [16; 45], making it well-suited for modeling complex data distributions. In other words, for any distribution $P_X(x)$, the normalizing flow f can convert it into a tractable distribution $P_Z(z)$. The whole generation process can be described mathematically as:

$$x = f(z) = f_n \odot f_{n-1} \odot \ldots \odot f_1(z) \tag{1}$$

The density estimation process can be expressed as:

$$z = f(x) = f_1^{-1} \odot f_2^{-1} \odot \ldots \odot f_n^{-1}(x) \tag{2}$$

where x represents the training sample, z represents the output of model f, n represents the number of cascaded layers, $\odot$ represents the cascading operation, and f^{-1} represents the inverse transformation.

The likelihood of the input data x can be estimated by a change of variables [40]:

$$P_X(x) = P_Z(z)|det\frac{\partial f^{-1}}{\partial x}| \tag{3}$$

$$logP_X(x) = logP_Z(z) + log|det\frac{\partial f^{-1}}{\partial x}| \tag{4}$$

where $|det\frac{\partial f^{-1}}{\partial x}|$ denotes the absolute determinant of the Jacobian matrix. Our goal is to maximize $P_X(x)$, which is equivalent to maximizing $logP_X(x)$ or minimizing $-logP_X(x)$. For simplicity, we usually set z to follow a Gaussian distribution $z \sim N(0, 1)$. Finally, we can get the following expression:

$$-logP_X(x) = \frac{||z||^2}{2} - log|det\frac{\partial f^{-1}}{\partial x}| \tag{5}$$

In the process of anomaly detection, we do not need the model to generate new samples. We only need a normalizing flow to map all normal samples into a simple distribution and then consider samples far away from this distribution as abnormal samples during testing. Therefore, we only need to use Eq (2) and (5) for training.

3.2. Overview of the Proposed Method

The overall framework of our model is shown in Fig. 2. Our model is a two-stage model. As can be seen from this figure, our model contains two important modules, CDCF and FSFlow. In stage 1, our main task is to achieve multimodal fusion. The specific method is to align the image features with the point-cloud features through the proposed **C**ross-**M**odal **A**lignment (**CMA**) network, and then fuse the aligned difference features (as compensation features) with the depth information. In stage 2, we use **F**requency-**S**patial **F**low (FSFlow) to map the features obtained in stage 1 to a Gaussian distribution. FSFlow consists of multiple flow blocks. In each block, we use the proposed FSE module to enhance the features, thus improving the anomaly detection performance.

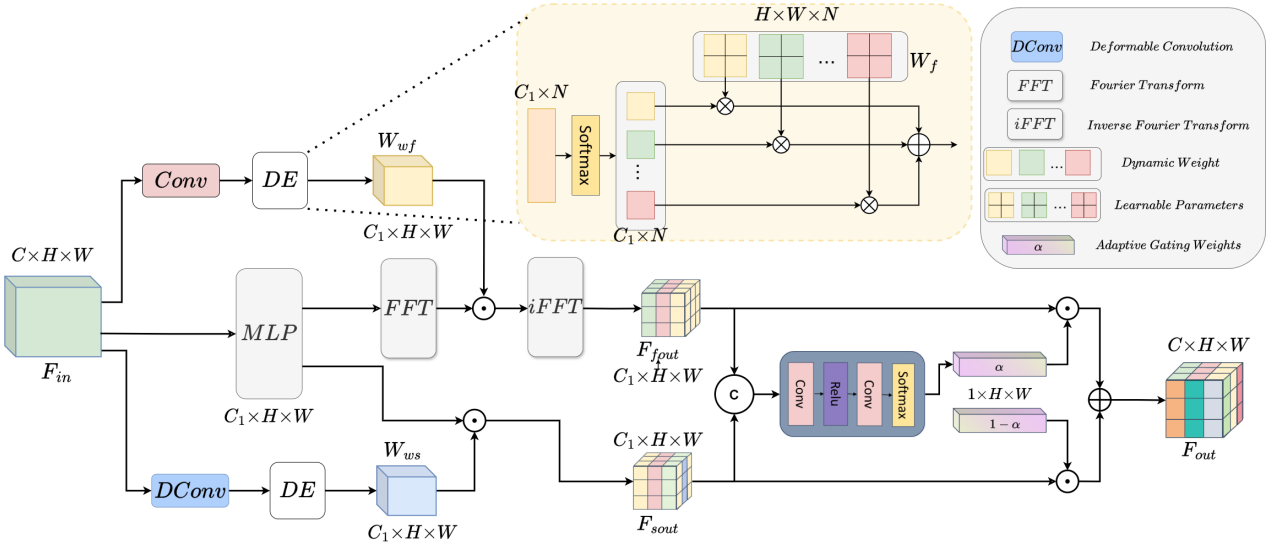

Figure 3: The detail of FSE, which models the features from both the frequency domain and the spatial domain and performs fusion of the features from the two domains through an adaptive gating mechanism to obtain the final features, where W_{wf} and W_{ws} are learnable weights, C and C_1 represent different channel dimensions, H and W represent the height and width of the feature, respectively, and N represents the number of layers.

3.3. Cross-modal Difference Compensation Fusion

When applying normalizing flow to multimodal anomaly detection, the first challenge is effectively fusing the features from different modalities. Following the approach in [56], we use feature extractors pre-trained on ImageNet [15] for images and on ShapeNet [6] for point clouds. However, there ared the fundamental differences between images and point clouds: images consist of structured 2D pixels that capture appearance and texture, while point clouds are made up of unordered 3D points that represent geometric structure. These differences in data representation and corresponding feature extractors lead to a semantic gap between cross-modal features. Simply concatenating or fusing features from both modalities often results in mismatched feature spaces. As a result, many existing multimodal anomaly detection methods, such as EasyNet [7], 3DSR [60], and CFM [10], adopt a late fusion strategy. However, late fusion operates at the output level and cannot effectively capture cross-modal interactions. In contrast, enabling interaction between modalities can significantly enhance model performance by leveraging complementary information [43; 11]. The foundation of effective multimodal interaction lies in achieving accurate cross-modal alignment.

To address this issue, we introduce a lightweight network called CMA (as illustrated in the right part of Fig. 2), inspired by recent multimodal anomaly detection methods such as M3DM [56] and CFM [10], which use two- and three-layer networks for feature alignment. Following this trend but different from the simple concatenation of aligned features in M3DM and CFM, we design an effective CMA to align the multimodal features and capture their differences. In addition, we employ an objective function that enforces alignment between the two modalities by applying cosine loss for global feature alignment and mean squared error (MSE) loss for local feature consistency, aiming to maximize the alignment of normal samples across modalities. Although the model structure is relatively simple, it effectively supports the subsequent integration of compensation features. During training, we input the point-cloud features F^i_{pc} of normal samples, and obtain the aligned features F^i_{pi} through the CMA module. We use the difference between F^i_{pi} and F^i_{img} as a loss function to guide the model to align these two modalities for normal samples.

Since the model is trained solely on normal samples, it does not learn the alignment relationships for abnormal samples in the test set. This leads to cross-modal discrepancies, or biases, which can serve as informative cues for representing anomalies. To leverage this, we incorporate these bias features into the model and fuse them with the depth map. In contrast to the baseline method asymmetric student-teacher (AST) network [48], which relies solely on depth information, our method utilizes richer and more informative features, making it more effective for anomaly detection tasks. We use the difference between these two features, i.e., F^i_{diff}, as a compensation for the 3D feature, and fuse it with the depth feature F^i_d to obtain the 3D feature F^i_{3D}. Then, we concatenate F^i_{3D} and F^i_{img}, before passing it to the normalizing flow for training as follows:

$$\begin{aligned} F^i_{fusion} &= concat(F^i_{img}, (F^i_{img} - CMA(F^i_{pc})) + F^i_d) \\ &= concat(F^i_{img}, F^i_{diff} + F^i_d) \\ &= concat(F^i_{img}, F^i_{3D}) \end{aligned} \tag{6}$$

where F^i_d represents the depth feature, as used in AST [48].

3.4. FSFlow

The normalizing flow in our model is shown in Fig. 2, which is composed of a series of flow blocks, following [16]. We denote the input and output of the i-th layer as F_i and

F_{i+1}, respectively, where F_1 is the output feature of the CDCF module in stage 1, which is equivalent to F_{fusion} in Eq. (6). Assuming that the dimension of F_i in channels is D, the model first randomly shuffles the input features and divides them into two features $F_{i,1:d}$ and $F_{i,d+1:D}$ along the channel dimension. Then, these two features are sent to the FSE module to obtain the scaling and translation parameters s_i^1 and t_i^1. Finally, we apply the parameters obtained, i.e., s_i^1 and t_i^1, to the features and concatenate the final results along the channel dimension. The above process is expressed in mathematical formulas as below:

$$F_{i,1:d}, F_{i,d+1:D} = split(F_i) \tag{7}$$

$$s_i^1, t_i^1 = FSE(F_{i,1:d}),$$
$$F_{i+1,d+1:D} = F_{i,d+1:D} \odot e^{s_i^1} + t_i^1 \tag{8}$$

$$s_i^2, t_i^2 = FSE(F_{i+1,d+1:D}),$$
$$F_{i+1,1:d} = F_{i,1:d} \odot e^{s_i^2} + t_i^2 \tag{9}$$

$$F_{i+1} = concat(F_{i+1,1:d}, F_{i+1,d+1:D}) \tag{10}$$

3.5. Frequency-Spatial Feature Enhancement

Due to domain gaps and local inconsistencies between modalities, fusing the multimodal features often introduces global and local noise. Existing multimodal anomaly detection methods struggle to effectively suppress such noise. To this end, features from both the frequency and spatial domains have been considered, such as [21; 62; 53]. However, these methods have mainly focused on single-modality data, without considering the local and global noise introduced by cross-modal fusion. In addition, existing normalizing flow-based methods [64; 31] rely solely on spatial information [64] or frequency domain information [31], but not both.

Inspired by GFilter [44], we adopt a strategy to learn frequency-specific filtering weights. GFilter performs global frequency filtering by multiplying the input with a learnable, input-independent external matrix. Building on this, DyGFilter [50] introduces dynamic filtering by generating weights conditioned on the input, allowing the filters to adapt based on data characteristics. Following this idea, we incorporate input-dependent dynamic filtering into our design.

Specifically, our network learns N dynamic frequency filters in parallel, where N is a hyperparameter that determines the number of external matrices used to generate a weighted composite filter, as shown in Fig. 3. In addition, we propose an approach to enhance feature representation by jointly modeling features from both the spatial and frequency domains. To address the irregular shapes commonly found in anomalous regions, we incorporate deformable convolution [12] in the spatial domain, enabling the network to adapt to varying object geometries and effectively capture key local regions within the input features.

We then apply FFT to convert the features into the frequency domain, as each frequency band carries distinct information that differentiates normal from abnormal images [33]. This transformation enables more effective global filtering by leveraging a broader receptive field, thus overcoming the local limitations of spatial convolution [44]. As demonstrated by GFilter [44], processing visual features in the frequency domain significantly improves global filtering efficiency. After transformation, we enhance the features by multiplying them with the learned frequency-domain weights.

The whole process can be described as follows:

$$W_{wf} = DE(Conv(F_{in})),$$
$$W_{ws} = DE(DConv(F_{in})) \tag{11}$$

$$F_{fout} = iFFT(FFT(MLP(F_{in})) \odot W_{wf}) \tag{12}$$

$$F_{sout} = MLP(F_{in}) \odot W_{ws} \tag{13}$$

$$F_{out} = Gate(F_{fout}, F_{sout}) \tag{14}$$

where F_{in} and F_{out} represent the input and output features of the FSE module, respectively, while W_{wf} and W_{ws} represent the learned weights. The dynamic enhancement (DE) module serves as the core component for feature dynamic enhancement. It introduces data dependencies [50] by applying convolution to the input and combining the result with a set of learnable parameters W_s or W_f. This enables the DE module to generate adaptive weights (i.e., W_{sf} and W_{wf}) based on real-time input, which are then used to enhance spatial and frequency features in the respective branches. The $MLP(\cdot)$ denotes a multi-layer perceptron operation. $Gate(\cdot)$ is a Gated Linear Unit (GLU) [13].

The feature fusion mechanism, shown in Fig. 3, first concatenates the frequency enhancement features and the spatial enhancement features in the channel dimension, and then passes it into the network to learn the final weights. The weights generated by the Softmax layer allow the model to select the most relevant features according to input data.

3.6. Loss Function

For CDCF, we use cosine similarity as the loss function, following [10]. To better align 3D point-cloud and 2D image features, we use the MSE loss as the loss function. The final loss of CDCF can be described as follows:

$$L_{CDCF} = \beta||F_{img}^i - CMA(F_{pc}^i)||_2 + (1 - \frac{F_{img}^i \cdot CMA(F_{pc}^i)}{||F_{img}^i|| \times ||CMA(F_{pc}^i)||}) \tag{15}$$

where β is a hyperparameter used to balance these two losses. For normalizing flow, following [16], we define the likelihood of F_{fusion} as below:

$$P_F(F_{fusion}) = P_Z(z)|det\frac{\partial z}{\partial F_{fusion}}| \tag{16}$$

where z is the output of the normalizing flow model. It maps the input features $F_{fusion} \in P_F(F_{fusion})$ to the latent features $z \in P_Z(z)$ through the normalizing flow. We can use the negative log-likelihood estimate to optimize the model. For ease of calculation, we assume that $z \sim N(0, 1)$. Finally, our loss function can be expressed as:

$$L_{nf} = -logP_F(F_{fusion})$$

Table 1
Anomaly detection result for multimodal anomaly detection (I-AUROC %) of MVTec 3D-AD. Optimal and sub-optimal I-AUROC results are in **bold** and underline, respectively.

Method	Source	Bagel	Cable Gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
DepthGAN [3]	Arxiv'21	53.8	37.2	58.0	60.3	43.0	53.4	64.2	60.1	44.3	57.7	53.2
DepthAE [3]	Arxiv'21	64.8	50.2	65.0	48.8	80.5	52.2	71.2	52.9	54.0	55.2	59.5
VoxelGAN [3]	Arxiv'21	68.0	32.4	56.5	39.9	49.7	48.2	56.6	57.9	60.1	48.2	51.7
VovelVM [3]	Arxiv'21	55.3	77.2	48.4	70.1	75.1	57.8	48.0	46.6	68.9	61.1	60.9
BTF [25]	CVPRW'23	91.8	74.8	96.7	88.3	93.2	58.2	89.6	91.2	92.1	88.6	86.5
AST [48]	WACV'23	98.3	87.3	97.6	97.1	93.2	88.5	97.4	**98.1**	**100**	79.7	93.7
EasyNet [7]	ACM MM'23	99.1	**99.8**	91.8	96.8	94.5	94.5	90.5	80.7	99.4	79.3	92.6
M3DM [56]	CVPR'23	99.4	90.9	97.2	97.6	96.0	94.2	97.3	89.9	97.2	85.0	94.5
ShapeGuided [9]	ICML'23	98.6	89.4	98.3	99.1	97.6	85.7	99.0	96.5	96.0	86.9	94.7
ITNM [55]	Nuecom'24	99.2	95.1	98.8	95.0	99.9	87.6	91.9	96.5	99.1	85.0	94.8
M3DM-NR [54]	Arxiv'24	99.3	91.1	97.7	97.6	96.0	92.2	97.3	89.9	95.5	88.2	94.5
LSFA [52]	ECCV'24	**100**	93.9	98.2	98.9	96.1	**95.1**	98.3	96.2	98.9	95.1	97.1
CFM [10]	CVPR'24	99.4	88.8	98.4	99.3	98.0	88.8	94.1	94.3	98.0	**95.3**	95.4
MMRD [20]	AAAI'24	99.9	94.3	96.4	94.3	99.2	91.2	94.9	90.1	99.4	90.1	95.0
3DSR [60]	WACV'24	98.1	86.7	**99.6**	98.1	**100**	**99.4**	98.6	97.8	**100**	**99.5**	**97.8**
Ours	-	**100**	98.0	97.5	**99.6**	98.7	93.4	**99.2**	97.3	**100**	92.5	97.6

$$= -(\log P_Z(z) + \log |det \frac{\partial z}{\partial F_{fusion}}|)$$
$$= \frac{||z||_2^2}{2} - \log |det \frac{\partial z}{\partial F_{fusion}}| \quad (17)$$

where $||z||_2^2$ denotes L2 norm and $|det \frac{\partial z}{\partial F_{fusion}}|$ denotes the absolute determinant of the Jacobian matrix.

4. Experiment

4.1. Datasets

We use two popular multimodal anomaly detection datasets: MVTec 3D-AD [3] and Eyecandies [5]. MVTec 3D-AD contains ten classes, a total of 2656 training samples, and 1137 test samples. The training set contains only normal samples, and the test set contains normal and abnormal samples. Each sample contains both the RGB image and the corresponding point-cloud. Eyecandies also contains ten classes. At the same time, the classes in the dataset involve a variety of lighting environments, which makes detection more challenging. There are 1000 training samples and 50 test samples for each class.

4.2. Implementation Details

Our proposed MADFlow method was implemented in the Pytorch framework, and all experiments were performed on a NVIDIA RTX4090 GPU.

The pretrained feature extractor we used is exactly the same as the one used by AST [48]. Additionally, we use PointMAE [39] as a point-cloud feature extractor, which is also the commonly used feature extractor for point-cloud data. In order to extract richer feature representations, we concatenate the outputs of the 19th, 26th, and 35th layers of the feature extractor and use the concatenated features as the final image features. For comparison, we use the experimental settings as in AST [48]. For FSFlow, we use 4 flow blocks that are conditioned on positional encoding with 32 channels. During training, the batch size is set to 8. For stage 1, the Adam optimizer [29] is used for training, using momentum parameters β_1=0.9 and β_2=0.999, a learning rate of 1e-5 and a weight decay of 1e-5 for training 50 epochs. For stage 2, the Adam optimizer [29] is used for training, using momentum parameters β_1=0.9 and β_2=0.999, but with a learning rate of 1e-3 and train for 200 epochs. For the hyperparameter N in FSE, we set it to 3. For the hyperparameter β of Eq. 15, we set it to 1.

4.3. Performance Metrics

For performance evaluations, we employ the widely used metrics for anomaly detection, such as I-AUROC and P-AUROC [41]. I-AUROC is the area under the Receiver Operating Characteristic (ROC) curve based on the overall prediction score of the sample, which is used to measure the model's ability to distinguish normal samples from abnormal samples. P-AUROC is the area under the ROC curve based on the pixel-level prediction results, which is used to evaluate the model's localization accuracy for abnormal areas, but P-AUROC may be affected by the size of the abnormal area.

4.4. Baseline Methods

To evaluate the performance of our method, we compared it with the previous SOTA methods for multimodal anomaly detection, including ShapeGuided [9], BTF [25], AST [48], EasyNet [7], M3DM [56], ITNM [55], M3DM-NR [54], CFM [10], LSFA [52], MMRD [20], and 3DSR [60].

4.5. Results

Table 1 describes the anomaly detection results of our method on the MVTec 3D-AD dataset. Our method achieves an average result of 97.6% on all classes, outperforming the previous SOTA method CFM [10] (**+2.2%**) and LSFA [52] (**+0.5%**). We noticed that there is still a gap between the results of our method and those of CFM and LSFA methods

Table 2
The results for multimodal anomaly detection (I-AUROC %) on the Eyecandies dataset. Optimal and sub-optimal I-AUROC results are in **bold** and underline, respectively.

Method	Candy Cane	Chocolate Cookie	Chocolate Praline	Confetto	Gummy Bear	Hazelnut Truffle	Licorice Sandwich	Lollipop	Marsh mallow	Peppermint Candy	Mean
BTF [25]	60.6	90.4	79.2	93.9	72.0	56.3	86.7	86.0	99.2	84.2	80.9
AST [48]	57.4	74.7	74.7	88.9	59.6	61.7	81.6	84.1	98.7	98.7	78.0
EasyNet [7]	73.7	93.4	86.6	96.6	71.7	82.2	84.7	86.3	97.7	96.0	86.9
M3DM [56]	62.4	95.8	**95.8**	**100**	88.6	75.8	94.9	83.6	**100**	**100**	89.7
ShapeGuided* [9]	50.2	90.7	77.8	93.4	74.4	61.4	82.4	87.9	98.4	99.0	81.6
ITNM [55]	-	-	-	-	-	-	-	-	-	-	-
LSFA [52]	-	-	-	-	-	-	-	-	-	-	-
CFM [10]	68.0	93.1	95.2	88.0	86.5	78.2	91.7	84.0	99.8	96.2	88.1
MMRD [20]	**85.4**	**100**	94.6	99.8	**90.8**	**94.7**	96.6	**98.4**	**100**	**100**	**94.0**
3DSR [60]	65.1	99.8	90.4	97.8	87.5	86.1	96.5	89.9	99.0	97.1	90.9
Ours	82.8	**100**	93.7	99.8	85.8	79.6	**97.4**	92.3	96.9	**100**	92.8

- indicates that no results have been released. * indicates that there are no released results, but we re-run the experiment using its official configuration.

in some classes. We attribute this to the fact that we only use 3D information as a compensation method to assist detection, and these classes are affected by the lighting environment, which makes image anomaly detection challenging.

Compared with other methods, our method ranks in the top three in eight out of ten classes, which also reflects the effectiveness of our fusion method. AST [48] also uses normalizing flow for multimodal anomaly detection, but only uses the depth information in the 3D information and uses another student network to enhance the detection effect of the model. Our method far exceeds its results, which further demonstrates the effectiveness of our proposed CDCF and FSE modules. BTF [25] uses the Fast Point Feature Histograms (FPFH) features as point-cloud features and shows that the features extracted using the pre-trained feature extractor [39] are not as effective as the FPFH features in clustering-based methods. This illustrates the complexity of point-cloud data and the large domain gap of the feature extractor. With the proposed CDCF, we can effectively alleviate the problem of the domain gap.

In addition, we visualize the detection effect of our model in Fig. 1. Specifically, compared to the baseline CFM [10], our anomaly score heatmap in the rightmost column of Fig. 1 shows a significant improvement in matching with the ground truth (GT) in the third column. We also visualized the final anomaly score distribution in Fig. 4. The red and blue bars represent the scores of abnormal and normal samples with respect to the learned distribution. A higher score indicates a greater deviation from the normal sample distribution and is therefore classified as abnormal. As shown in the figure, our method significantly improves the scores of anomalous classes, enabling more precise anomaly detection.

To further demonstrate the effectiveness of our model, we also compare it with baseline methods on the Eyecandies dataset [5]. The results are shown in Table 2. Note that some baselines in Table 1 do not appear in Table 2 because they have no results released on this Eyecandies dataset. Although Eyecandies is a very challenging dataset, our method still achieves very competitive results. For the Eyecandies dataset, which contains samples with complex lighting variations, RGB images are often affected by reflections and highlights. Our proposed CDCF module depends on the alignment

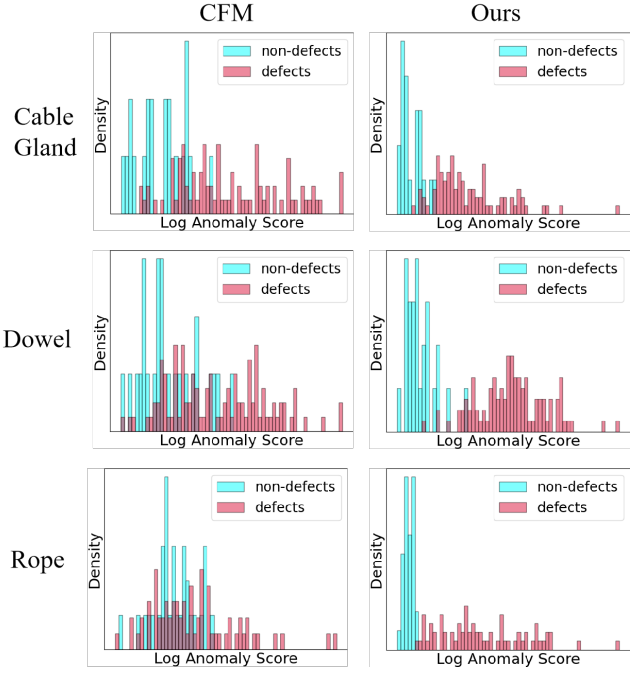

Figure 4: Comparison of anomaly score distribution of some classes in the MVTec 3D-AD [3] dataset. The first column shows the results of CFM [10]. The second column shows our results, which demonstrate the improved separation between abnormal and normal samples, with minimal overlap between blue and red, reflecting the better anomaly detection performance of our proposed method.

between RGB images and point clouds, but under challenging lighting conditions, RGB images can introduce noise, leading to inaccurate cross-modal alignment and ineffective compensation fusion. In contrast, MMRD [20] primarily utilizes depth maps, which are less sensitive to lighting, making it more robust in such scenarios. In future work, we plan to explore illumination-invariant techniques to address this limitation. In Fig. 5, we visualize the detection effects of our method and CFM [10] on the Eyecandies dataset, respectively. In the visualization results, darker red regions indicate a higher likelihood of anomalies, while blue areas suggest lower anomaly scores. By comparing these highlighted regions with the ground truth, we can better assess the detection

Table 3
Anomaly detection result for multimodal anomaly detection (I-AUROC % and P-AUROC %) of MVTec 3D-AD and Eyecandies. **Average** represents the average result of two datasets. Optimal and sub-optimal results are in **bold** and underline, respectively.

Method	MVTec 3D-AD		Eyecandies		Average	
	I-AUROC	P-AUROC	I-AUROC	P-AUROC	I-AUROC	P-AUROC
BTF	86.5	99.2	80.9	-	83.7	-
AST	93.7	97.6	78.0	90.2	85.9	93.9
EasyNet	92.6	91.9	86.9	-	89.8	-
M3DM	94.5	99.2	89.7	97.7	92.1	98.5
ShapeGuided	94.7	**99.6**	81.6	95.3	88.2	97.5
ITNM	94.8	99.5	-	-	-	-
M3DM-NR	94.5	98.9	-	-	-	-
LSFA	97.1	99.3	-	-	-	-
CFM	95.4	99.3	88.1	97.4	91.8	98.4
MMRD	95.0	99.2	**94.0**	**98.3**	94.5	**98.8**
3DSR	**97.8**	99.5	90.9	-	94.4	-
3DSR*	91.4	96.7	87.9	88.8	89.7	92.8
Ours	97.6	98.6	92.8	96.9	**95.2**	97.8

- indicates that there is no reported results and no released source code.
* indicates that we rerun the code under the consistent setting with most of baselines (such as M3DM, MMRD).

performance by the model, demonstrating that our method achieves highly competitive results.

Comparison of different evaluation metrics. We report the performance of our method in comparison with other approaches using the positioning metrics P-AUROC, as shown in Table 3. The methods BTF [25], M3DM [56], M3DM-NR [54], ShapeGuided [9], ITNM [55], and LSFA [52] are all memory-based anomaly detection techniques. These approaches operate by comparing the input features with the normal features stored in a memory bank, selecting the most similar entries to compute the anomaly score. This strategy mitigates the uncertainty associated with global reconstruction errors and generally yields high P-AUROC scores. However, the primary objective of anomaly detection is to determine whether a sample is abnormal, and in this aspect, our method demonstrates superior performance in terms of I-AUROC. On the other hand, EasyNet [7], CFM [10], and 3DSR [60] represent reconstruction-based methods. EasyNet separately reconstructs RGB and depth images, merging the outputs via late fusion, a process that neglects cross-modal interactions. In contrast, CFM explicitly learns cross-modal relationships through matching mechanisms, enabling richer feature representation. Reconstruction-based approaches typically rely on pixel-wise optimization during training, which improves their ability to recover fine details [49]. Although our method is somewhat less effective at detecting fine-grained local anomalies, it excels at capturing global semantic features. In general, considering the average performance across both datasets, our method achieves the highest I-AUROC score of 95.2, indicating superior robustness compared to baseline methods for identifying abnormal samples in diverse scenarios.

Complexity Analysis. Table 4 compares the number of parameters and inference time of our method with other approaches. Among non flow-based methods, BTF [25], M3DM [56], and Shape-Guided [9] are memory-based and require comparing test features with all stored normal features during inference, resulting in longer processing times. Although CFM [10] uses only three simple linear layers, leading to low parameter count and FLOPs, our method still outperforms it in detection accuracy by +2.2% I-AUROC. In contrast, 3DSR [60] involves multiple networks, leading to significantly higher computational cost. However, as it avoids using pre-trained feature extractors, it offers better real-time performance. In contrast, our method attains higher detection accuracy with only a moderate increase in parameters and FLOPs, making it well-suited for industrial anomaly detection scenarios where high accuracy is essential. For flow-based approaches, AST [48] employs both a teacher and a student network, leading to a higher parameter count and computational load. Although AST avoids using point clouds, our method incorporates an additional step of extracting point cloud features via PointMAE, which introduces extra computational overhead and results in a lower Frames Per Second (FPS) compared to AST.

Table 4
Complexity of the proposed method as compared with baseline methods.

	Methods	Params(M)	FLOPs(G)	FPS
Non Flow-based	BTF	-	-	1.2
	EasyNet	-	-	-
	M3DM	15.8	12.4	0.5
	ShapeGuided	3.2	1.3	1.5
	ITNM	-	-	-
	M3DM-NR	-	-	-
	LSFA	-	-	-
	CFM	5.5	3.2	12.6
	MMRD	-	-	-
	3DSR	58.7	261.3	21.9
Flow-based	AST	86.5	49.8	18.9
	Ours	27.4	15.7	7.1

- indicates that there are no reported results, or source code for reproducing the results.

4.6. Ablation Study

We conduct a series of ablation experiments to verify the effectiveness of our proposed modules. It is worth noting

Table 5
Ablation studies on each module. The best results are in **bold**

CDCF	FSE	I-AUROC	P-AUROC
×	×	95.5	96.6
✓	×	96.7	97.8
×	✓	96.8	97.5
✓	✓	**97.6**	**98.6**

Table 6
Ablation studies on mutilmodal fusion strategy. The best result are in **bold**.

Methods	I-AUROC	P-AUROC
*Concat**	70.3	96.5
CA+	96.3	97.5
CDCF (cosine loss)	97.4	97.9
CDCF (MSE loss)	97.2	98.2
CDCF (KL loss)	96.5	97.6
CDCF (contrasive loss)	97.1	98.0
CDCF (KL+MSE)	97.0	98.2
CDCF (KL+cosine)	97.2	98.1
CDCF (contrastive+MSE)	96.8	97.9
CDCF (contrastive+cosine)	97.3	98.1
CDCF	**97.6**	**98.6**

* indicates that the image features and point cloud features are directly concatenated in the channel dimension.
+ indicates directly adding the image features and the point cloud features after the CMA network.

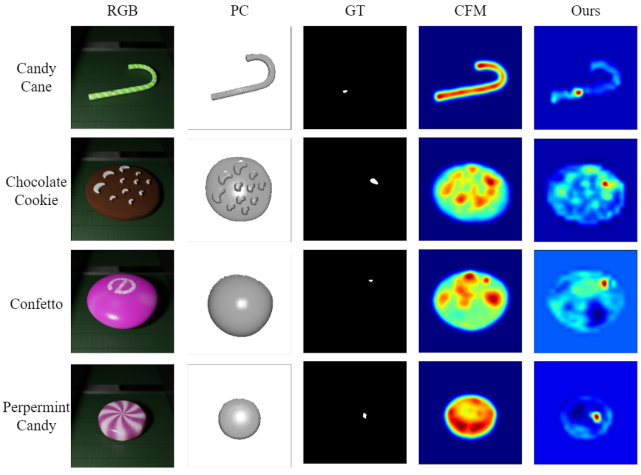

Figure 5: Illustration of Eyecandies datasets [3]. The first and second column represent the input RGB image and point-cloud, respectively. The third column represents the ground truth. The fourth column is the anomaly detection heat map of the latest competitive method CFM [10], and the fifth column is the heat map of our proposed method.

that all of our ablation experiments are performed on the MVTec 3D-AD dataset. In addition to anomaly identification (I-AUROC), which is the most important indicator for evaluating anomaly detection models, we provide the anomaly localization (P-AUROC) indicator to further study the model's anomaly localization ability.

Effectiveness of proposed components. First, we use AST [48] as our baseline and keep the feature extraction consistent with our method. The experimental results are shown in Table 5. As can be seen from the results in Table 5, the CDCF module significantly improves the performance of the model due to the use of the structural information brought by the point clouds. At the same time, FSE further improves the ability to capture anomalies by modeling from both frequency and spatial perspectives.

Comparison among different fusion methods. The results for different fusion methods are shown in Table 6. The concat indicates that the image features and point cloud features are directly concatenated in the channel dimension, where the image features and point-cloud features are extracted by the pre-trained feature extractor without any alignment operation. From the results, we can see that directly concatenating the two features at the channel dimension achieves poor results. This is probably because there is a large domain gap in the representation of the data of the two modalities, making it difficult to map the concatenated features to a simple distribution.

The CA results show that the model performs better when using fused features compared to image features alone; however, it is still outperformed by our proposed CDCF. This is likely because CA merely appends point-cloud features to image features, incorporating only a limited amount of 3D information. While this addition is somewhat robust to environmental noise, it does not fully exploit the valuable information available in point clouds as effectively as our CDCF does. However, when image features are degraded by environmental factors, the combined representation remains partially influenced by these disturbances, limiting the effectiveness of the fusion. Therefore, the results of CA are better than those of *Concat*, but still worse than those of our proposed CDCF. Our proposed CDCF can mitigate this influence by utilizing the learned compensation information between modalities.

Comparison among different loss functions. We further compared the effects of the loss functions on the results. From Table 6, we can see that the best performance was achieved using cosine loss and MSE. The two loss functions we employed are complementary and help balance differences across various scenarios. Specifically, the cosine loss emphasizes global differences in high-dimensional features while remaining insensitive to absolute value discrepancies. In contrast, the MSE loss is sensitive to exact numerical differences, capturing local variations across individual feature dimensions [63]. Thus, combining these two loss functions allows the model to balance both global high-dimensional feature alignment and local variations across feature dimensions, thereby enhancing the stability of the overall optimization process. We also performed an ablation experiment on the impact of the hyperparameter β in the loss function on the results in Table 7.

Although KL divergence and contrastive loss only achieved suboptimal results, they still achieved significant performance improvements compared to the method based on direct feature concatenation.

Table 7
Ablation studies of the hyperparameter β on the MVTec 3D-AD dataset. The best results are in **blod**.

Hyperparameter	MVTec 3D-AD		Eyecandies	
	I-AUROC	P-AUROC	I-AUROC	P-AUROC
β=0.3	97.1	97.7	92.1	96.5
β=0.5	97.3	98.1	92.4	96.6
β=0.8	97.4	98.3	**92.8**	**96.9**
β=1	**97.6**	**98.6**	92.6	96.8
β=1.2	97.5	98.3	92.3	96.6
β=1.5	97.4	98.4	92.1	96.5
β=2	97.2	98.5	91.7	95.8

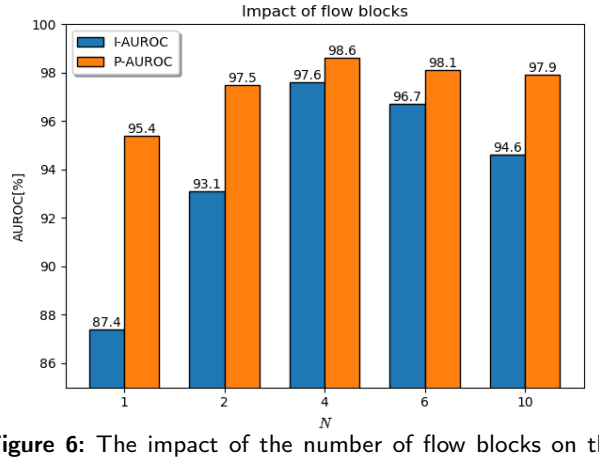

Figure 6: The impact of the number of flow blocks on the detection results. The horizontal axis represents the number of flow blocks, and the vertical axis represents the overall performance (AUROC %).

Table 8
Ablation studies on frequency-spatial enhancement. The best results are in **blod**.

Methods	I-AUROC	P-AUROC
$Frequency^{+}$	97.2	98.2
$Spatial^{*}$	96.9	98.1
SE [26]	96.4	97.7
CBAM [57]	96.2	97.3
$FSE(add)^{-}$	97.4	98.4
FSE (w/o DE)	97.2	98.3
FSE (w/o DConv)	97.4	98.5
FSE	**97.6**	**98.6**

+ denotes that only frequency enhancement in FSE is used.
* denotes that only spatial enhancement in FSE is used.
$^{-}$ denotes directly adding the frequency enhancement features and spatial enhancement feature.

Comparison among different feature enhancement methods. Table 8 shows the results of improving feature representation from different perspectives. It can be seen that the proposed FSE, due to the use of both frequency and spatial information, achieves better results than using only the frequency or spatial information. In addition, we compared our method with the commonly used attention mechanism. From the results shown in Table 8, we can see that our method offers better performance on the I-AUROC and P-AUROC metrics than the attention mechanism.

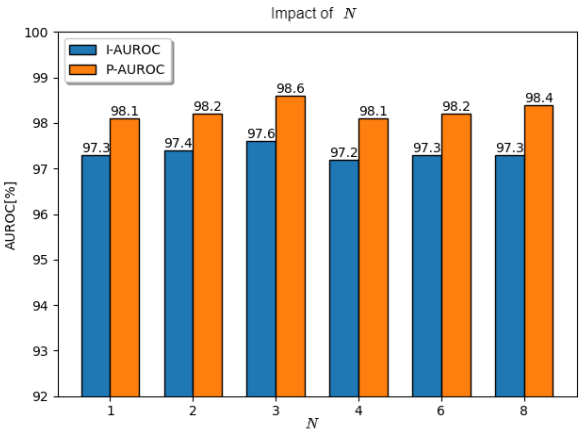

Figure 7: The impact of hyperparameter N on the detection results. The horizontal axis represents the number of flow blocks, and the vertical axis represents the overall performance (AUROC %).

Table 9
Ablation studies of hyperparameter $C1$. The best results are in **blod**.

$C1 = nC$	I-AUROC	P-AUROC
$C1 = 0.5C$	97.51	98.59
$C1 = 0.75C$	97.53	98.60
$C1 = C$	**97.58**	**98.62**
$C1 = 2C$	97.47	98.55
$C1 = 3C$	97.49	98.59

The impact of the number of flow blocks. We studied the impacts of different number of flow blocks on the final detection results. It can be seen from Fig. 6 that when the number of blocks is less than 4, the detection accuracy increases with increasing number of blocks. This is because when the number of blocks is small, it is difficult for the model to accurately map the complex distribution to a specified Gaussian distribution. However, as the number of blocks continues to increase, this model involves more layers, resulting in a higher computational load.

The impact of hyperparameters N and C_1. We conducted an ablation experiment on the hyperparameter N to study its impact on performance. The experimental results are shown in Fig. 7. It can be seen from the figure that the performance of the model tends to be relatively stable with respect to the setting of N, with $N = 3$ giving better results than the other options. As shown in Table 9, if C_1 is too large, the model may introduce redundant information, including noise, during the learning process. However, our method is robust to the choice of C_1.

5. Conclusion

In this paper, we have studied the use of normalizing flow to process image and point-cloud data simultaneously, and proposed MADFlow for multimodal anomaly detection. To make better use of point-cloud features, we propose CDCF and use it to compensate for depth features. Finally, we proposed FSE to model features from frequency and spatial

perspectives simultaneously, which alleviated the uncertainty in anomaly types and demonstrated competitive performance, as compared with the state of the art baselines on commonly used multimodal anomaly detection datasets. However, the proposed method also has limitations. For example, since our model relies primarily on image data, combined with compensation from the CDCF module, its performance will degrade when the image modality is corrupted, e.g. by the lighting condition. In future work, we can incorporate lighting-invariant strategies and strengthen the representation of geometric information.

References

[1] Bao, L., Zhou, X., Zheng, B., Cong, R., Yin, H., Zhang, J., Yan, C., 2025. Ifenet: Interaction, fusion, and enhancement network for vdt salient object detection. IEEE Transactions on Image Processing .

[2] Bergmann, P., Fauser, M., Sattlegger, D., Steger, C., 2019. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pp. 9592–9600.

[3] Bergmann, P., Jin, X., Sattlegger, D., Steger, C., 2021. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. arXiv preprint arXiv:2112.09045 .

[4] Bi, C., Li, Y., Luo, H., 2024. Dual-branch reconstruction network for industrial anomaly detection with rgb-d data, in: International Conference on Image, Signal Processing, and Pattern Recognition, SPIE. pp. 767–774.

[5] Bonfiglioli, L., Toschi, M., Silvestri, D., Fioraio, N., De Gregorio, D., 2022. The eyecandies dataset for unsupervised multimodal anomaly detection and localization, in: Proceedings of the Asian Conference on Computer Vision, pp. 3586–3602.

[6] Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., et al., 2015. Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 .

[7] Chen, R., Xie, G., Liu, J., Wang, J., Luo, Z., et al., 2023. Easynet: An easy network for 3d industrial anomaly detection, in: Proceedings of the 31st ACM International Conference on Multimedia, pp. 7038–7046.

[8] Cho, M., Kim, T., Kim, W.J., Cho, S., Lee, S., 2022. Unsupervised video anomaly detection via normalizing flows with implicit latent features. Pattern Recognition 129, 108703.

[9] Chu, Y.M., Chieh, L., Hsieh, T.I., Chen, H.T., Liu, T.L., 2023. Shape-guided dual-memory learning for 3d anomaly detection .

[10] Costanzino, A., Ramirez, P.Z., Lisanti, G., Di Stefano, L., 2024. Multimodal industrial anomaly detection by crossmodal feature mapping, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17234–17243.

[11] Cui, Y., Chen, R., Chu, W., Chen, L., Tian, D., Li, Y., Cao, D., 2021. Deep learning for image and point cloud fusion in autonomous driving: A review. IEEE Transactions on Intelligent Transportation Systems 23, 722–739.

[12] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., et al., 2017. Deformable convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 764–773.

[13] Dauphin, Y.N., Fan, A., Auli, M., Grangier, D., 2017. Language modeling with gated convolutional networks, in: Proceedings of the 34th International Conference on Machine Learning, pp. 933–941.

[14] Deng, H., Li, X., 2022. Anomaly detection via reverse distillation from one-class embedding, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pp. 9737–9746.

[15] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., et al., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on Computer Vision and Pattern Recognition, Ieee. pp. 248–255.

[16] Dinh, L., Sohl-Dickstein, J., Bengio, S., 2016. Density estimation using real nvp. arXiv preprint arXiv:1605.08803 .

[17] Fang, H., Zhang, T., Zhou, X., Zhang, X., 2024. Learning better video query with sam for video instance segmentation. IEEE Transactions on Circuits and Systems for Video Technology .

[18] Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I., 2023. Imagebind: One embedding space to bind them all, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 15180–15190.

[19] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Communications of the ACM 63, 139–144.

[20] Gu, Z., Zhang, J., Liu, L., Chen, X., Peng, J., et al., 2024. Rethinking reverse distillation for multi-modal anomaly detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8445–8453.

[21] Guo, Z., Chen, H., He, F., 2024. Msfnet: Multi-scale spatial-frequency feature fusion network for remote sensing change detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing .

[22] He, H., Dong, X., Zhou, X., Wang, B., Zhang, J., 2024. Interactive fusion and correlation network for three-modal images few-shot semantic segmentation. IEEE Signal Processing Letters 31, 2430–2434.

[23] Hirschorn, O., Avidan, S., 2023. Normalizing flows for human pose anomaly detection, in: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13499–13508.

[24] Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Advances in neural information processing systems 33, 6840–6851.

[25] Horwitz, E., Hoshen, Y., 2023. 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.

[26] Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 7132–7141.

[27] Kim, D., Baik, S., Kim, T.H., 2023. Sanflow: Semantic-aware normalizing flow for anomaly detection. Advances in Neural Information Processing Systems 36, 75434–75454.

[28] Kingma, D.P., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 .

[29] Kingma, D.P., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

[30] Kirichenko, P., Izmailov, P., Wilson, A.G., 2020. Why normalizing flows fail to detect out-of-distribution data. Advances in Neural Information Processing Systems 33, 20578–20589.

[31] Lei, J., Hu, X., Wang, Y., Liu, D., 2023. Pyramidflow: High-resolution defect contrastive localization using pyramid normalizing flow, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pp. 14143–14152.

[32] Li, C.L., Sohn, K., Yoon, J., Pfister, T., 2021. Cutpaste: Self-supervised learning for anomaly detection and localization, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pp. 9664–9674.

[33] Liang, Y., Zhang, J., Zhao, S., Wu, R., Liu, Y., et al., 2023. Omni-frequency channel-selection representations for unsupervised anomaly detection. IEEE Transactions on Image Processing .

[34] Liu, G., Lan, S., Zhang, T., Huang, W., Wang, W., 2021. Sagan: skip-attention gan for anomaly detection, in: 2021 IEEE International Conference on Image Processing, IEEE. pp. 2468–2472.

[35] Liu, T., Li, B., Du, X., Jiang, B., Geng, L., Wang, F., Zhao, Z., 2023. Fair: Frequency-aware image restoration for industrial visual anomaly detection. arXiv preprint arXiv:2309.07068 .

[36] Lv, C., Zhou, X., Wan, B., Wang, S., Sun, Y., Zhang, J., Yan, C., 2024. Transformer-based cross-modal integration network for rgb-t salient object detection. IEEE Transactions on Consumer Electronics 70, 4741–4755.

[37] Ma, W., Lan, S., Huang, W., Wang, W., Yang, H., et al., 2023. A semantics-aware normalizing flow model for anomaly detection, in: 2023 IEEE International Conference on Multimedia and Expo (ICME),

pp. 2207–2212.
[38] Ma, W., Li, Y., Lan, S., Wang, W., Huang, W., et al., 2024. Semantic-aware normalizing flow with feature fusion for image anomaly detection. Neurocomputing 590, 127728.
[39] Pang, Y., Wang, W., Tay, F.E., Liu, W., Tian, Y., et al., 2022. Masked autoencoders for point cloud self-supervised learning, in: European Conference on Computer Vision, Springer. pp. 604–621.
[40] Papamakarios, G., Nalisnick, E., Rezende, D.J., Mohamed, S., Lakshminarayanan, B., 2021. Normalizing flows for probabilistic modeling and inference. Journal of Machine Learning Research 22, 1–64.
[41] Powers, D.M., 2020. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061 .
[42] Pumarola, A., Popov, S., Moreno-Noguer, F., Ferrari, V., 2020. C-flow: Conditional generative flow models for images and 3d point clouds, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7949–7958.
[43] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR. pp. 8748–8763.
[44] Rao, Y., Zhao, W., Zhu, Z., Lu, J., Zhou, J., 2021. Global filter networks for image classification. Advances in Neural Information Processing Systems 34, 980–993.
[45] Rezende, D., Mohamed, S., 2015. Variational inference with normalizing flows, in: International conference on machine learning, PMLR. pp. 1530–1538.
[46] Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., et al., 2022. Towards total recall in industrial anomaly detection, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pp. 14318–14328.
[47] Rudolph, M., Wehrbein, T., Rosenhahn, B., Wandt, B., 2022. Fully convolutional cross-scale-flows for image-based defect detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1088–1097.
[48] Rudolph, M., Wehrbein, T., Rosenhahn, B., Wandt, B., 2023. Asymmetric student-teacher networks for industrial anomaly detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2592–2602.
[49] Salehi, M., Mirzaei, H., Hendrycks, D., Li, Y., Rohban, M.H., Sabokrou, M., 2021. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. arXiv preprint arXiv:2110.14051 .
[50] Tatsunami, Y., Taki, M., 2024. Fft-based dynamic token mixer for vision, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 15328–15336.
[51] Tien, T.D., Nguyen, A.T., Tran, N.H., Huy, T.D., Duong, S., Nguyen, C.D.T., Truong, S.Q., 2023. Revisiting reverse distillation for anomaly detection, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pp. 24511–24520.
[52] Tu, Y., Zhang, B., Liu, L., Li, Y., Zhang, J., et al., 2025. Self-supervised feature adaptation for 3d industrial anomaly detection, in: European Conference on Computer Vision, Springer. pp. 75–91.
[53] Wang, C., Jiang, J., Zhong, Z., Liu, X., 2023a. Spatial-frequency mutual learning for face super-resolution, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 22356–22366.
[54] Wang, C., Zhu, H., Peng, J., Wang, Y., Yi, R., et al., 2024a. M3dm-nr: Rgb-3d noisy-resistant industrial anomaly detection via multimodal denoising. arXiv preprint arXiv:2406.02263 .
[55] Wang, J., Wang, X., Hao, R., Yin, H., Huang, B., et al., 2024b. Incremental template neighborhood matching for 3d anomaly detection. Neurocomputing 581, 127483.
[56] Wang, Y., Peng, J., Zhang, J., Yi, R., Wang, Y., et al., 2023b. Multimodal industrial anomaly detection via hybrid fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8032–8041.
[57] Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, pp. 3–19.
[58] Yang, G., Huang, X., Hao, Z., Liu, M.Y., Belongie, S., Hariharan, B., 2019. Pointflow: 3d point cloud generation with continuous normalizing flows, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 4541–4550.
[59] Yao, H., Liu, M., Wang, H., Yin, Z., Yan, Z., Hong, X., Zuo, W., 2024. Glad: Towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection. arXiv preprint arXiv:2406.07487 .
[60] Zavrtanik, V., Kristan, M., Skočaj, D., 2024. Cheating depth: Enhancing 3d surface anomaly detection via depth simulation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2164–2172.
[61] Zhang, X., Li, S., Li, X., Huang, P., Shan, J., et al., 2023. Destseg: Segmentation guided denoising student-teacher for anomaly detection, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pp. 3914–3923.
[62] Zhang, X., Xu, C., Fan, G., Hua, Z., Li, J., Zhou, J., 2025. Fscmf: A dual-branch frequency-spatial joint perception cross-modality network for visible and infrared image fusion. Neurocomputing , 130376.
[63] Zhao, H., Gallo, O., Frosio, I., Kautz, J., 2016. Loss functions for image restoration with neural networks. IEEE Transactions on computational imaging 3, 47–57.
[64] Zhou, Y., Xu, X., Song, J., Shen, F., Shen, H.T., 2025. MSFlow: Multiscale flow-based framework for unsupervised anomaly detection. IEEE Transactions on Neural Networks and Learning Systems 36, 2437–2450.

A. Appendix

A.1. False detection analysis

We also present failure cases in Fig. 8, where our model was unable to correctly identify anomalies. The first column shows the RGB image, the second displays the point cloud, the third visualizes the model's predicted anomaly heatmap, and the fourth shows the ground truth. In these cases, our method successfully detected one anomaly but missed small defects at the bottom, i.e. the areas that are also challenging for manual inspection. In the third row, although the model exhibits increased attention to abnormal regions, the signal is not strong enough to classify them as anomalies, suggesting limitations in handling complex depth variations. In the fourth row, strong lighting interference significantly degrades the RGB image, which is our primary modality in the CDCF fusion framework. As a result, the performance of the model is compromised. In future work, we plan to incorporate lighting invariance and strengthen geometric feature representation to better handle such scenarios.

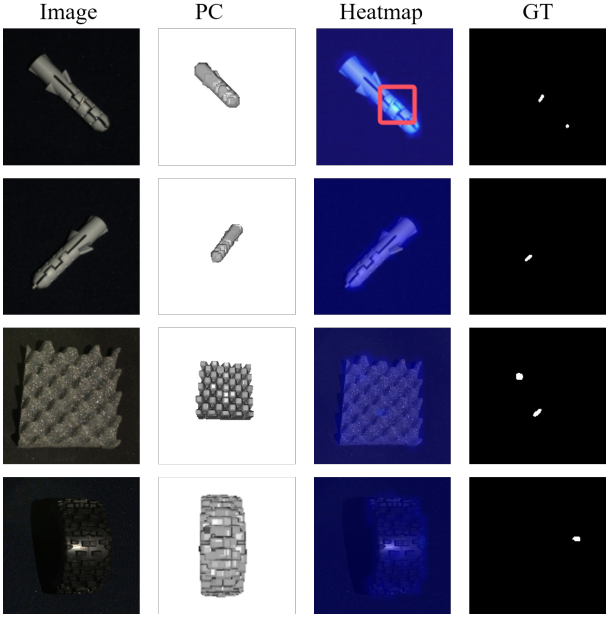

Figure 8: The first and second columns represent the input image and point cloud. The third column is the heat map produced by our model. The fourth column represents the actual abnormal location.

A.2. Anomalies in different modalities

We illustrate scenarios where anomalies appear in only one or both modalities in Fig. 9. The first row shows an anomaly present only in the RGB image, where the point cloud fails to capture the color-related irregularity. The second row depicts an anomaly visible solely in the point cloud where the RGB image is obscured due to lighting conditions, while the point cloud remains unaffected by illumination. The third row presents a case where the anomaly is visible in both the image and the point cloud. From the visualizations, it is evident that our method effectively handles all three types of scenarios.

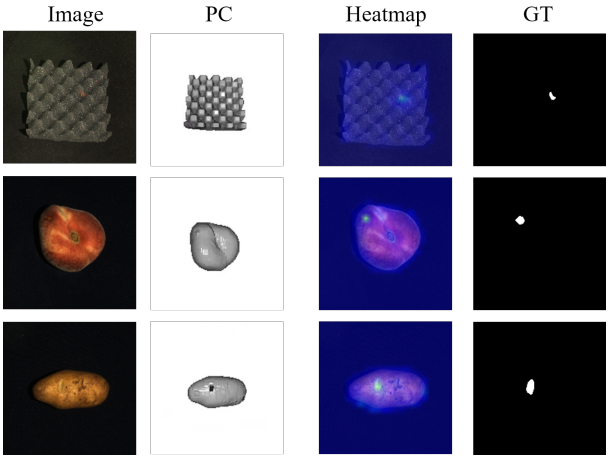

Figure 9: The first and second columns represent the input image and point cloud. The third column is the heat map produced by our model. The fourth column represents the actual abnormal location.

Table 10
Ablation studies of CMA layer number on the MVTec 3D-AD dataset. The best results are in **blod**.

Layers	I-AUROC	P-AUROC
1	96.5	97.8
2	96.7	97.9
3	97.2	98.3
4	97.6	98.6
6	**97.6**	**98.5**

Table 11
Ablation studies of different layers of feature extractor on the MVTec 3D-AD dataset. The best results are in **blod**.

Layers	I-AUROC	P-AUROC
35	97.0	97.9
26+35	97.4	98.2
19+26+35	97.62	98.6

A.3. Number of layers in the CMA network

We conducted a comparison of different CMA layer depths, as shown in Table 10. When increasing the number of layers from 1 to 4, both I-AUROC and P-AUROC improved consistently, indicating that deeper architectures significantly enhance cross-modal feature alignment. With too few layers, the model struggles to learn complex mapping relationships, i.e. shallow convolutions fail to capture global structural dependencies, leading to suboptimal alignment and limited anomaly detection performance. However, when the number of layers exceeds 4, the model tends to overfit to noise in the training data rather than learning generalizable features, which ultimately reduces accuracy and adds unnecessary computational overhead.

A.4. The impact of different layers of feature extractor

We evaluated how the number of layers in the feature extractor affects detection performance. As shown in Table 11,

utilizing multiple layers simultaneously leads to improved results. Although deeper layers typically capture higher-level semantic features, anomaly detection also heavily relies on low-level structural features. To account for this, we incorporate features from shallower layers during feature selection and include them in the final image feature representation through concatenation.