IMMoE: Incomplete Multi-View Anomaly Detection via Mixture of View Experts Fusion

Lei Hu

South China University of Technology, Guangzhou, China
hulei777@foxmail.com

Abstract. Existing Multi-view Anomaly Detection (MAD) methods assume that all views are completely available and model each view separately. However, in real industrial scenarios, information in the view may be missing due to faults such as occlusion, which leads to the performance degradation of existing methods due to the lack of a multi-view consistency prior. To address this, we explored a more challenging task: Incomplete Multi-View Anomaly Detection (IMVAD), in which some areas of each view were masked. We proposed a pipeline for automatically generating the IMVAD dataset and generated the **RIMAD** dataset based on the Real-IAD dataset through this pipeline. In addition, in order to effectively utilize the information of multiple views in the absence of view information, we propose **IMMoE**, which consists of two key modules: (1) Multi-View Expert Fusion (MVEF) effectively fuses multi-view information through a multi-view expert network and guides the reconstruction of a single view; (2) Local Anomaly Enhancement Encoder (LAEE) effectively prevents the model from overfitting the mask region by applying dropout to local features. Our method achieves state-of-the-art performance on both the RIMAD and Real-IAD datasets, especially on RIMAD, we have increased the pixel-level and image-level metrics by 11.8% and 2.8%, respectively. Our source code is available at https://github.com/HULEI7/IMMoE.

Keywords: Anomaly Detection · Incomplete Multi-View · Mixture of Experts

1 Introduction

Industrial Anomaly Detection (IAD) has significantly enhanced production efficiency and product quality by automatically identifying product defects [1–3, 13, 15, 21, 27]. However, traditional IAD methods primarily rely on single-view images, which inevitably suffer from blind spots and limited surface coverage. To ensure comprehensive inspection of complex industrial components, the task of Multi-view Anomaly Detection (MAD) has emerged [5, 37], aiming to utilize information from multiple viewpoints for a holistic evaluation.

However, current mainstream MAD methods [8, 10, 23] usually assume that every view is complete and model each view separately. Essentially, this approach

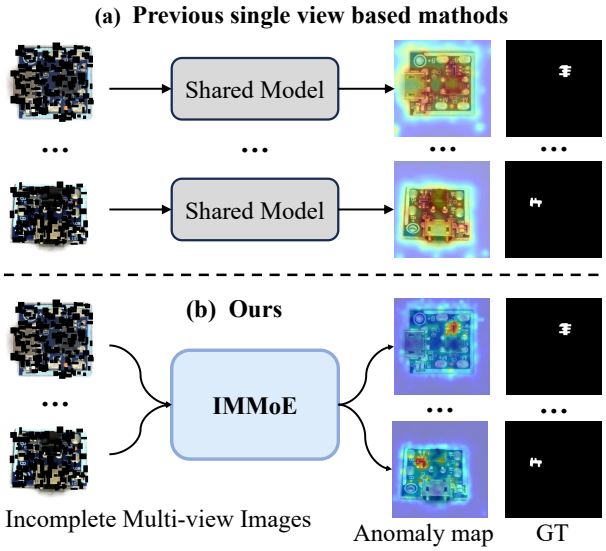

Fig. 1: Comparisons of our method with the previous single-view-based method on IMVAD. (a) Due to the lack of multi-view interaction, previous methods will experience significant performance degradation when partial information is missing. (b) Our method effectively integrates the useful information among various views through a multi-view expert network, thereby achieving effective information complementarity. See Figure 4 for more comparisons.

is still just single-view anomaly detection, and it fails to bring out the true potential of multi-view datasets. In real-world industrial scenarios, problems like occlusion or equipment errors often lead to missing information in different views. As a result, the performance of existing methods drops sharply because they lack the necessary correlation between these views, as shown in Figure 1.

To tackle these limitations, we propose a more challenging task: **I**ncomplete **M**ulti-**V**iew **A**nomaly **D**etection (**IMVAD**). In this task, parts of each view are masked, and the central challenge is to leverage information from other views to reconstruct the missing regions within the current view. To support this task, we proposed an automatic pipeline for generating the IMVAD dataset, as shown in Figure 2. Specifically, we first employ Segment Anything (SAM) [17] to extract the product masks, after which we simulate view incompleteness by randomly discarding image content within selected regions. Following this procedure, we constructed the **RIMAD** dataset based on Real-IAD [37]. In this dataset, 50% of the target areas in each view are masked to reflect missing information.

As illustrated in Figure 1 and Table 1, we evaluated the performance of current state-of-the-art methods on the RIMAD dataset. Our observations reveal a substantial degradation in their performance, which can be attributed to two primary factors: (1) **The absence of cross-view consistency**, existing methods often treat each view independently and lack a mechanism to ensure consis-

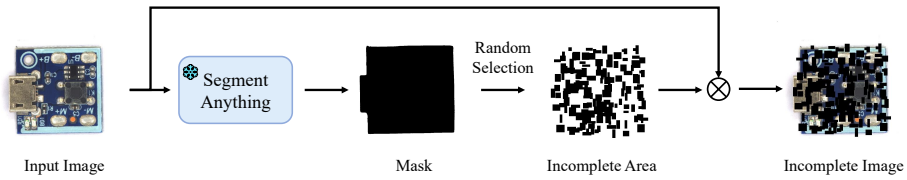

Fig. 2: The overall production process of RIMAD. We first employ Segment Anything to extract the product masks, after which we simulate view incompleteness by randomly discarding image content within selected regions.

tency between different views. When a view is incomplete, the model cannot borrow useful information from other views to fill in the gaps, leading to unreliable detection results. (2) **Overfitting of the incomplete area**, Conventional models are prone to overfitting to the specific patterns of missing data. Instead of learning robust features, the network may treat the masked regions as inherent data characteristics. This leads to a reduction in generalization, where the model struggles to differentiate between genuine structural anomalies and artifacts caused by data incompleteness.

To tackle these limitations, as illustrated in Figure 3, we introduce **IMMoE**, a framework comprising two synergistic modules: (1) Multi-View Expert Fusion (**MVEF**), which leverages a multi-view expert network to effectively aggregate cross-view information, thereby providing robust guidance for single-view reconstruction and ensuring inter-view consistency. (2) Local Anomaly Enhancement Encoder (**LAEE**), which mitigates overfitting to masked regions by employing a localized dropout mechanism on feature maps. This regularizes the model to learn more discriminative representations and enhances its sensitivity to genuine anomalies. Furthermore, to enable the model to perceive the varying learning difficulties across different areas, we introduce a **Area Adaptive Loss**. Specifically, we partition the image into three distinct categories: incomplete product regions, complete product regions, and non-product (background) areas. By assigning different gradient weights to these three zones, we allow the model to adaptively focus on the most challenging parts, ensuring a more effective optimization process based on the inherent difficulty of each region. Extensive experiments show that our model achieves the best performance on both RIMAD and Real-IAD [37] datasets, outperforming other existing methods.

To sum up, our contributions are as follows:

- We identify and define a realistic and challenging task: Incomplete Multi-View Anomaly Detection (**IMVAD**). To support this research, we develop an automated generation pipeline and construct **RIMAD**, the first benchmark dataset specifically designed to evaluate model robustness under view incompleteness scenarios.
- We propose **IMMoE**, a novel framework tailored for the IMVAD task. By leveraging cross-view correlations and region-adaptive learning strategies,

it effectively restores multi-view consistency and prevents the model from overfitting to masked regions.
– Extensive experiments demonstrate that our method achieves state-of-the-art performance on both the constructed RIMAD and the public Real-IAD datasets, especially on RIMAD, we have increased the pixel-level and image-level metrics by 11.8% and 2.8%, respectively.

2 Related Work

2.1 Multi-view Anomaly Detection

Existing MAD methods can be broadly categorized into two types: **single-view-based** approaches and **multi-view fusion-based** approaches.

Single-view-based methods treat each view as an independent input and model them separately. UniAD [39] proposes a unified Transformer-based framework to handle multiple categories within a single network. SimpleNet [22] achieves efficient localization by injecting noise into pre-trained features for discriminative training. MambaAD [10] leverages State Space Models to enhance computational efficiency and capture long-range dependencies. Dinomaly [8] advocates a minimalist philosophy, demonstrating that pre-trained features require only minimal adaptation. INP-Former [23] focuses on extracting intrinsic normal prototypes directly from individual images as robust references.

Multi-view fusion-based approaches aim to integrate information across multiple perspectives. IDIF [24] introduces an intra-view decoupling and inter-view fusion strategy to better capture cross-view correlations. MVAD [12] introduces the MVAS algorithm, which utilizes a window-based attention mechanism to aggregate features from the most semantically correlated regions across multiple views. However, despite the cross-view interactions incorporated by these methods, they remain susceptible to overfitting on the incomplete regions when facing the IMVAD task, ultimately leading to a significant performance degradation.

2.2 Mixture of Experts

The Mixture of Experts (MoE) paradigm has emerged as a powerful architecture for scaling model capacity while maintaining computational efficiency through conditional computing [7, 18, 25, 32, 33]. The core mechanism relies on a gating network that dynamically routes inputs to a sparse subset of specialized experts [16]. Recent advancements have significantly extended the stability and efficiency of this routing process, such as the single-expert routing in Switch Transformers [6] and the balanced competitive learning in V-MoE [31]. MoEAD [26] utilizes a Mixture-of-Experts pool within a recursive Transformer block. By adaptively selecting class-specific experts via a gating mechanism, it achieves parameter-efficient multi-class anomaly detection, balancing model compact-ness with high discriminative performance across diverse industrial categories. PromptMoE [32] addresses zero-shot anomaly detection by replacing

monolithic prompt learning with a compositional approach. It utilizes a pool of expert prompts (semantic primitives) and a visually-guided Mixture-of-Experts (MoE) mechanism. An image-gated router dynamically aggregates these experts to generate instance-adaptive textual representations, achieving superior generalization across diverse, unseen anomaly categories. Unlike scaling-oriented MoE, we reframe IMMoE for multi-view fusion. It leverages specialized experts to aggregate complementary information from all perspectives into a unified global representation. This global context then interacts with individual views, enabling the model to effectively restore incomplete regions by borrowing knowledge from the integrated multi-view features.

3 IMVAD

3.1 Problem Definition

In real-world industrial scenarios, information loss in product images is often inevitable due to equipment malfunctions or physical occlusions. To evaluate model robustness under such realistic constraints, we propose a challenging task: Incomplete Multi-View Anomaly Detection (IMVAD). Specifically, the input for IMVAD in each viewpoint i consists of two components: the incomplete view image x_i and its corresponding binary mask m_i, the mask m_i explicitly indicates the missing regions, where $m_i(p) = 0$ denotes an incomplete pixel and $m_i(p) = 1$ represents a valid pixel at position p. The model is required to reconstruct the original normal image even when the input view information is incomplete.

3.2 RIMAD Dataset

As illustrated in Figure 2, to construct the RIMAD dataset, we simulate realistic view incompleteness through a structured masking pipeline. We first employ Segment Anything (SAM) [17] to precisely extract the product masks, ensuring that the simulation is focused solely on the object of interest. To simulate information loss, we randomly discard image content within the segmented product regions. Specifically, we populate the product area with randomly distributed rectangles, each with an area ranging from 8 to 32 square pixels. This process is iteratively applied until **50%** of the total product area is obscured. This granular and stochastic masking strategy creates a challenging environment that forces the model to rely on cross-view correlations to reconstruct the missing industrial details.

4 Methods

4.1 Overview

The overall framework of our method is illustrated in Figure 3. Similar to INP-Former [23] and Dinomaly [8], we adopt a feature reconstruction-based frame-

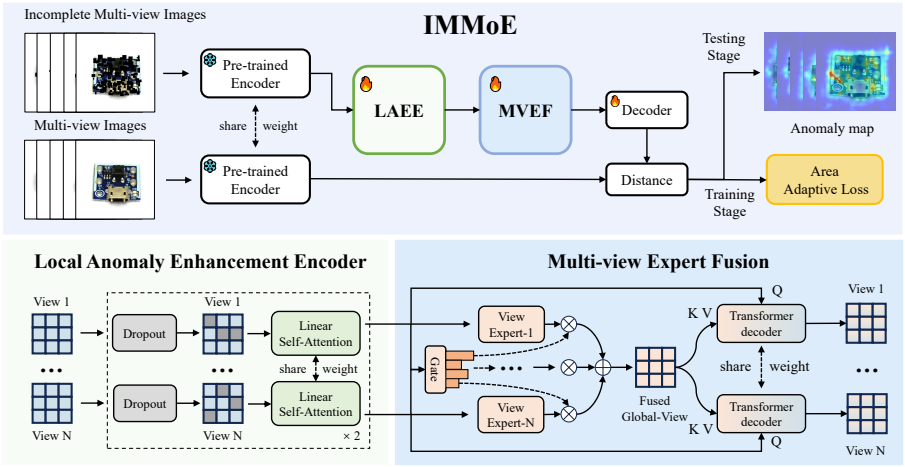

Fig. 3: The overall architecture of the IMMoE framework. Our reconstruction-based pipeline centers on two synergistic modules: the **Local Anomaly Enhancement Encoder** (LAEE), which prevents overfitting to masked regions by simulating anomaly patterns through stochastic feature discarding, and the **Multi-view Expert Fusion** (MVEF), which adaptively aggregates cross-view complementary information via a Mixture-of-Experts mechanism. Together, they ensure robust feature restoration and inter-view consistency, enabling precise anomaly localization even under severe information deficiency.

work. The overall framework consists of four key components and their corresponding intermediate features. First, in the **Feature Encoding** stage, a pre-trained encoder extracts representations from both the incomplete and original images, denoted as F_{in} and F_{gt}, respectively. Next, F_{in} is fed into the **Local Anomaly Enhancement Encoder** to generate enhanced features F_{enh} by randomly discarding feature elements to simulate potential anomaly patterns. Subsequently, the **Multi-view Expert Fusion** module integrates these enhanced features across all perspectives, producing a fused representation F_{fuse} that captures complementary information from multiple views. Finally, the **Reconstruction Decoder** utilizes F_{fuse} to generate the restored features F_{rec}. During training, the model is optimized by minimizing the cosine distance between F_{rec} and F_{gt} using a area-adaptive loss, while in the inference phase, the anomaly map is derived directly from the discrepancy between these two feature sets.

4.2 Local Anomaly Enhancement Encoder

The primary motivation for the Local Anomaly Enhancement Encoder is to prevent the model from overfitting to the incomplete areas. Conventional reconstruction models are often prone to overfitting the specific patterns of missing data; instead of learning robust semantic features, the network may inadver-

tently treat the masked regions as inherent data characteristics. This leads to a significant reduction in generalization, where the model struggles to differentiate between genuine structural anomalies and artificial artifacts caused by data incompleteness. Inspired by Dinomaly [8] and Dropout [35], we incorporate a stochastic feature-discarding mechanism to prevent the model from overfitting.

Specifically, for the incomplete view features $F_{in}^{i} \in \mathbb{R}^{N \times C}$, where i denotes the i-th view of the sample, N represents the number of image patches, and C represents the feature dimension, we first employ a dropout mechanism to randomly discard feature elements:

$$F_{drop}^{i} = Dropout(F_{in}^{i}, p), \tag{1}$$

where p denotes the dropout ratio.

Following the dropout operation, we incorporate a Linear Attention [8, 9, 34] mechanism to process the features of each view. The primary motivation is to recapture the spatial dependencies that may be disrupted by the stochastic discarding of features. By leveraging the global receptive field of attention, the model can effectively compensate for the missing information by aggregating context from the remaining active patches. We employ linear attention instead of standard self-attention to maintain computational efficiency while ensuring that each view's representation is sufficiently refined and robust before being passed to the subsequent fusion stage.

Specifically, the formula for Linear Attention (LA) is as follows:

$$Q = F_{drop}^{i} W^{Q}, K = F_{drop}^{i} W^{K}, V = F_{drop}^{i} W^{V}, \tag{2}$$

$$F_{enh}^{i} = LA(Q, K, V) = \sigma(Q)(\sigma(K^{T})V), \tag{3}$$

where W^{Q}, W^{K}, W^{V} are learnable parameters, $\sigma(x)$ is activation function.

Finally, the abnormally enhanced features of each view F_{enh}^{i} will be further sent to the multi-view expert fusion network for global view fusion.

4.3 Multi-view Expert Fusion

The motivation for the Multi-view Expert Fusion (MVEF) module stems from the inherent information deficiency of individual views. While the LAEE enhances the robustness of features within a single perspective, it cannot inherently recover semantic information that is completely lost due to occlusion or masking. In industrial anomaly detection, different viewpoints often provide complementary geometric and structural information. For instance, a defect obscured in one view may be clearly visible or inferable from another.

To fully exploit this multi-view synergy, the MVEF module is designed to dynamically aggregate cross-view knowledge. Instead of a simple concatenation or

averaging, we employ a Mixture-of-Experts (MoE) architecture. This approach allows the model to adaptively select and weight information from different views based on their relevance to the missing regions. By establishing a global context through these specialized experts, the MVEF provides the necessary cross-view guidance for each individual view, ensuring that the subsequent reconstruction is informed by a comprehensive understanding of the object's overall geometry rather than isolated, incomplete data.

Specifically, for the set of enhanced multi-view features $\{F_{enh}^i\}_{i=1}^V$ belonging to a single sample, we first concatenate them along the view dimension to serve as the input for the gating network:

$$F_{enh} = cat(F_{enh}^1; F_{enh}^2; \cdots; F_{enh}^V) \in \mathbb{R}^{V \times N \times C}, \tag{4}$$

where V represents the total number of views.

The gating network is composed of a linear layer and a softmax function, which are utilized to calculate the importance scores for each individual view:

$$F_{gate} = softmax(linear(F_{enh})) \in \mathbb{R}^{V \times N}, \tag{5}$$

After obtaining the importance scores for each view, we employ a dedicated view-specific expert to encode each view independently. Each expert is constructed with two linear layers and an activation function, designed to extract deep semantic features from its corresponding perspective:

$$F_{expert}^i = linear_1^i(gelu(linear_2^i(F_{enh}^i))) \in \mathbb{R}^{N \times C}, \tag{6}$$

After obtaining the output from each view-specific expert, we compute the weighted sum by multiplying each expert's feature map by its corresponding importance score. These weighted features are then aggregated to produce the final global fused representation F_{fuse}:

$$F_{fuse} = \sum_{i=1}^V F_{expert}^i F_{gate}^i, F_{fuse} \in \mathbb{R}^{N \times C}, \tag{7}$$

Upon obtaining the global fused representation F_{fuse}, we employ a cross-attention [36] to integrate this global information back into each individual view:

$$Q = F_{enh}^i W^Q, K = F_{fuse} W^K, V = F_{fuse} W^V, \tag{8}$$

$$F_{enh}^i = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V, \tag{9}$$

Following the integration of global information, we employ a transformer-based [14, 19, 20, 28, 29, 36] decoder utilizing linear attention to generate the final reconstructed features F_{rec}^i.

4.4 Area Adaptive Loss

To address the imbalance in reconstruction difficulty, inspired by INP-Former [23] and Dinomaly [8], we propose the Area Adaptive Loss (AAL), which dynamically modulates gradient magnitudes to prioritize informative regions over simple background textures. Specifically, we scale the loss by the ratio of local loss to total batch loss and incorporate region-specific priors that assign higher weights to incomplete areas and product foregrounds. This strategy ensures the optimization focuses on task-relevant and information-deficient regions, preventing the model from being overwhelmed by trivial background details.

To minimize the discrepancy between the original and reconstructed features, we first define the **cosine distance** as:

$$D_{cos}(a, b) = 1 - \frac{a^T \cdot b}{\|a\|\|b\|} \tag{10}$$

The basic reconstruction loss for the i-th view is thus $\mathcal{L} = D_{cos}(F^i_{gt}, F^i_{rec})$. We modulate the gradient of each patch j by an adaptive weight $w(j)$, formulated as:

$$\hat{f}(j) = cg(f(j)) \cdot w(j), \quad j \in [1, N] \tag{11}$$

The weighting factor is defined as:

$$w(j) = (\frac{D_{cos}(F^{i,j}_{gt}, F^{i,j}_{rec})}{\mu(D_{cos})})^\theta + \beta \tag{12}$$

where the first term emphasizes challenging patches by calculating the ratio of local loss to the batch mean $\mu(D_{cos})$. The second term, β, serves as a region-prior to incorporate spatial importance: we assign $\beta = 0.4$ for incomplete (masked) regions, $\beta = 0.2$ for product foregrounds, and 0 otherwise. This dual-weighting mechanism ensures that the optimization prioritizes task-relevant and information-deficient areas over trivial background textures.

5 Experiments

5.1 Experiments Setup

Datasets. We conducted our experiments on the two datasets, RIMAD and Real-IAD [37]. The Real-IAD contains samples of a total of 30 categories, and each sample contains images of 5 views. There were a total of 99,721 normal sample images, of which 36,465 were used for training and the remaining 63,256 for testing. There were 51,329 abnormal sample images, all of which were used for testing. We obtained the RIMAD dataset by masking 50% of the area in each view of the Real-IAD dataset. The specific method is shown in Figure 2.

Compared Methods. We compared with the most advanced multi-view anomaly detection methods currently available, including: UniAD [39], DeSTSeg [42], SimpleNet [22], DiAD [11], MambaAD [10], ViTAD [41], Dinomaly [8], INP-Former [23], MVAD [12].

Table 1: Comparison with state-of-the-art methods on **RIMAD**. The results are presented in Image-AUROC/Pixel-AP. **Bold** indicates the best performance, while underline indicates the second-best performance. All methods are reproduced under a unified experimental setting for fair comparison. †: method designed for Multi-view Anomaly Detection. More results are presented in the supplementary materials.

Method →	MambaAD	ViTAD	Dinomaly	INP-Former	MVAD†	**Ours†**
Category ↓	NeurIPS 24	CVIU 25	CVPR 25	CVPR 25	TMM 26	
Audiojack	81.7/32.8	78.0/21.5	83.1/34.5	77.1/27.8	76.7/26.9	**86.2**/**38.3**
Bottle Cap	**92.4**/**37.1**	83.4/23.4	82.5/35.5	84.6/30.1	92.1/36.8	86.7/34.0
Button Battery	64.8/20.7	65.6/39.6	74.0/39.7	74.9/**47.4**	62.3/23.4	**82.4**/32.9
End Cap	73.0/9.0	68.7/3.4	71.8/9.7	67.9/4.8	71.9/8.0	**74.4**/**10.2**
Eraser	**89.0**/33.0	88.9/30.6	83.9/30.3	88.3/**35.9**	88.8/32.5	85.0/30.4
Fire Hood	82.7/34.4	77.8/24.3	71.0/13.1	**85.4**/**35.0**	79.9/33.7	83.4/27.6
Mint	65.3/8.7	68.6/12.7	61.5/4.7	**71.8**/**13.7**	66.4/7.8	71.4/11.9
Mounts	**82.9**/**33.1**	81.4/25.5	76.8/23.3	79.0/23.6	82.6/33.0	78.7/23.4
PCB	81.7/28.8	64.7/11.7	62.7/11.8	70.8/13.1	70.0/18.3	**86.1**/**45.2**
Phone Battery	85.4/20.8	78.8/16.4	85.6/50.1	84.8/31.4	79.9/17.9	**87.8**/**58.8**
Plastic Nut	86.5/34.4	81.4/21.6	77.3/12.9	**88.4**/**36.0**	85.8/32.9	88.1/33.6
Plastic Plug	79.2/**28.2**	71.4/11.0	47.5/1.9	**83.8**/25.0	77.8/26.8	80.8/20.7
Porcelain Doll	83.3/27.5	80.2/22.2	82.8/18.2	**88.7**/24.6	84.1/**28.2**	85.1/19.2
Regulator	43.8/0.5	45.2/0.4	60.8/9.6	60.6/5.1	42.4/0.3	**70.6**/**20.4**
Rolled Strip Base	92.6/30.3	96.1/29.0	94.6/48.6	93.1/36.7	93.3/29.7	**96.4**/**49.6**
SIM Card Set	92.6/41.8	92.7/40.2	94.4/44.8	94.9/44.6	87.6/**49.3**	**95.2**/44.4
Switch	85.4/34.8	78.4/32.5	77.0/26.7	82.9/35.5	78.7/22.7	**94.4**/**59.4**
Tape	**95.6**/**43.1**	93.2/34.2	89.5/**43.1**	93.7/42.6	95.0/43.0	91.5/42.6
Terminal Block	**93.9**/40.1	83.4/21.5	90.1/48.0	88.8/40.5	90.1/36.3	92.6/**51.0**
Toothbrush	**83.2**/27.8	82.9/**30.1**	70.0/12.6	80.8/19.4	79.3/24.3	82.1/22.0
Toy	63.9/5.4	66.7/7.0	70.2/9.5	70.3/9.2	56.2/4.8	**74.0**/**12.3**
Toy Brick	81.9/**27.2**	75.3/23.9	70.4/14.6	74.7/17.9	**82.0**/26.6	72.2/14.2
Transistor1	**92.0**/37.3	84.1/32.3	87.9/**44.9**	88.1/42.9	87.1/32.6	89.3/**44.9**
U Block	77.0/24.6	77.6/**28.0**	82.1/20.9	82.6/26.1	77.0/23.6	**85.0**/25.8
USB	82.4/36.5	77.2/21.1	87.2/39.1	81.6/21.7	69.5/21.9	**91.6**/**42.5**
USB Adaptor	82.4/24.3	76.0/13.3	81.3/**24.8**	**83.4**/21.5	80.9/23.1	82.7/23.0
Vcpill	79.6/43.0	85.2/**55.5**	65.4/21.2	**89.7**/53.8	79.4/39.8	87.0/50.1
Wooden Beads	**85.6**/30.0	80.0/31.4	80.6/26.8	84.7/**32.4**	84.9/30.1	82.9/26.6
Woodstick	79.0/40.4	72.1/37.3	76.7/36.9	79.4/**42.3**	79.1/40.1	**80.1**/34.3
Zipper	97.8/32.2	97.6/**56.0**	92.8/54.9	98.3/47.1	97.6/48.4	**98.6**/45.1
Average	81.9/28.9	78.4/25.2	77.7/27.1	82.4/29.6	79.3/27.4	**84.7**/**33.1**

Metrics. Consistent with previous methods, we evaluate image-level anomaly classification using Area Under the Receiver Operating Characteristics (AUROC) [43]. For pixel-level anomaly segmentation, we adopt Area Under Precision-Recall (AP) [40]. We prioritize AP for segmentation because anomaly regions are significantly smaller than normal areas, AP provides a more reliable assessment by focusing on the minority class in such highly imbalanced pixel distributions.

Implementation Details. We adopt ViT-Base/14 with DINO2-R [4] as the pre-trained image encoder. The input image is first resized to 448^2 and then center-cropped to 392^2. We employ the StableAdamW optimizer [38] equipped with AMSGrad [30] to train the network for 50,000 iterations. The mask ratio of the RIMAD dataset has been set to 50%. We set the learning rate to 2e-3, β coefficients to $(0.9, 0.999)$, and weight decay to 1e-4. We set the dropout ratio

Table 2: Comparison with state-of-the-art methods on **Real-IAD**. **Bold** indicates the best performance, while underline indicates the second-best performance. †: method designed for Multi-view Anomaly Detection. All methods are reproduced under a unified experimental setting for fair comparison.

Method	Venue	I-AUROC	P-AP
UniAD	NeurIPS 22	83.0	21.1
DeSTSeg	CVPR 23	82.3	37.9
SimpleNet	CVPR 23	71.1	11.5
DiAD	AAAI 24	75.6	2.9
MambaAD	NeurIPS 24	86.3	33.0
ViTAD	CVIU 25	82.7	24.3
Dinomaly	CVPR 25	<u>89.3</u>	42.8
INP-Former	CVPR 25	89.1	<u>43.7</u>
MVAD†	TMM 26	86.6	30.3
Ours†		**89.7**	**44.3**

p to 0.1, and θ to 3. All the experiments were conducted on a single NVIDIA RTX 4090 GPU.

5.2 Comparison with Sota Methods

Quantitative Evaluation. As demonstrated in Table 1 and Table 2, our IMMoE framework achieves state-of-the-art (SOTA) performance across both RIMAD and Real-IAD datasets.

On the RIMAD dataset (Table 1), which specifically simulates view incompleteness, our method reaches 84.7% I-AUROC and 33.1% P-AP, outperforming the SOTA single-view method INP-Former [23] by 2.8% and 11.8%, respectively. Crucially, compared to the specialized multi-view method MVAD [12], our approach shows a significant margin of 5.4% in I-AUROC and 5.7% in P-AP. This gap reveals that while MVAD struggles to aggregate reliable features from corrupted inputs (79.3% I-AUROC), our MVEF module effectively reconstructs missing semantic priors by synergizing information across viewpoints.

On the Real-IAD dataset (Table 2), where all five views are complete, our method maintains its lead with 89.7% I-AUROC and 44.3% P-AP. It is noteworthy that while many recent high-performing methods like Dinomaly [8] and INP-Former [23] show strong results on complete data, our framework still surpasses them, further widening the gap over MVAD [12] (86.6% I-AUROC and 30.3% P-AP). This consistent superiority across both datasets underscores that our view-expert fusion strategy is not merely a remedy for missing data, but a robust architecture capable of adaptively capturing complex cross-view correlations regardless of information density.

Qualitative Evaluation. Figure 4 presents the heatmap visualizations of the anomaly regions predicted by our proposed method alongside other state-of-the-art techniques. It is evident that existing methods frequently struggle with

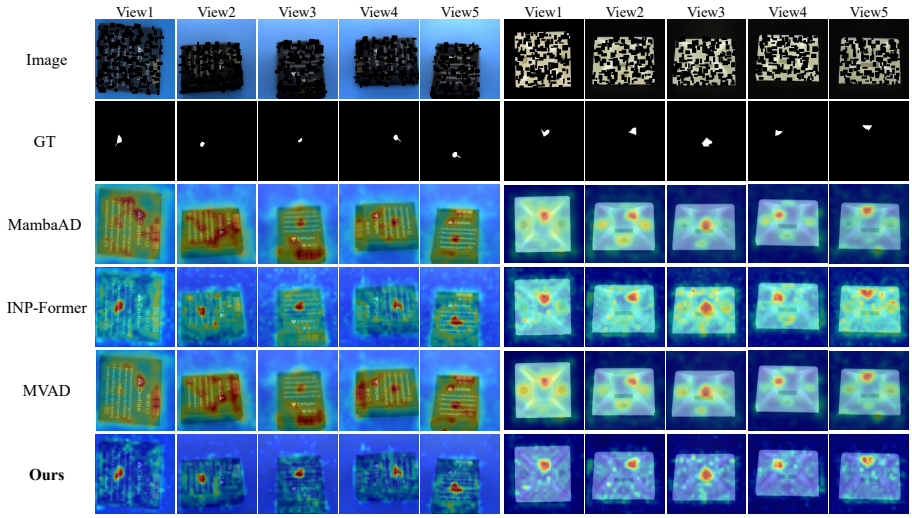

Fig. 4: Visualization of anomalous region segmentation. Areas with high anomaly scores are shown in red, while areas with low anomaly scores are shown in blue.

Table 3: Ablations of model design on RIMAD. LAEE: Local Anomaly Enhancement Encoder. MVEF: Multi-view Expert Fusion. AAL: Area Adaptive Loss.

LAEE	MVEF	AAL	I-AUROC	P-AP
✗	✗	✗	76.9	26.2
✓	✗	✗	81.3	29.1
✓	✓	✗	<u>84.0</u>	<u>32.3</u>
✓	✓	✓	**84.7**	**33.1**

information deficiency in incomplete views, often resulting in large-scale prediction errors, such as misidentifying normal background textures as anomalies or producing fragmented, noisy localization maps. In sharp contrast, by effectively integrating synergistic multi-view information, our approach suppresses such background interference and yields significantly more precise and comprehensive anomaly localizations. Our heatmaps demonstrate superior spatial consistency and align more closely with the actual defect boundaries, confirming that our model's cross-view reconstruction capability is crucial for high-fidelity anomaly segmentation even under severe occlusions.

5.3 Ablation Study

Model Design. To evaluate the contribution of each component, we conduct an ablation study on the RIMAD dataset (Table 3). Integrating the Local Anomaly Enhancement Encoder (LAEE) raises the Image-AUROC from 76.9% to 81.3%,

Table 4: Ablations of mask ratio on RIMAD (default=50%).

Ratio	Dinomaly	INP-Former	Ours
0%	89.3 / 42.8	89.1 / 43.7	**89.7** / **44.3**
25%	76.5 / 21.6	77.1 / **22.1**	**77.9** / 20.0
50%	77.7 / 27.1	82.4 / 29.6	**84.7** / **33.1**
75%	84.4 / 39.8	83.7 / 34.1	**87.3** / **42.0**

Table 5: Ablations of dropout rate on RIMAD (default=0.1).

Dropout rate	I-AUROC	P-AP
0.0	83.0	29.6
0.1	**84.7**	33.1
0.2	84.2	**33.4**
0.3	76.9	26.8

confirming its ability to prevent overfitting through feature dropout. The addition of Multi-view Expert Fusion (MVEF) further boosts performance to 84.0% I-AUROC, demonstrating that cross-view synergy is vital for recovering lost semantic details. Finally, the Area Adaptive Loss (AAL) yields the best results (84.7% I-AUROC, 33.1% P-AP), proving that dynamic gradient modulation effectively prioritizes challenging, task-relevant regions. These consistent gains validate the complementary nature of our synergistic modules.

Mask Ratio. As illustrated in Table 4 and Figure 5, we investigate the model's sensitivity to different mask ratios on RIMAD. An intriguing observation is that performance is relatively lower at moderate mask ratios (e.g., 25%) but improves as the ratio increases to 75%. We attribute this to a shift in the model's learning objective: at lower mask ratios, the model tends to rely heavily on the remaining local spatial features of the current view, which may be ambiguous or contain misleading cues from the incomplete boundaries. In contrast, at higher mask ratios, the model is "forced" to abandon local dependencies and instead functions as a pure reconstruction task, leveraging robust cross-view semantic priors and global context. This transition reduces the interference from corrupted local information, allowing the Multi-view Expert Fusion (MVEF) to more effectively aggregate reliable information from other views, thereby leading to superior and more stable anomaly detection performance.

Dropout Rate. We investigate the sensitivity of the LAEE to the dropout rate p, as shown in Table 5. The performance peaks at a dropout rate of 0.1, demonstrating that a moderate level of feature discarding effectively regularizes the model and prevents overfitting to masked regions. When the dropout rate is too low (0.0), the model lacks sufficient regularization, whereas a rate that is too high (0.3) leads to an excessive loss of critical semantic information, causing a sharp drop in performance. This parabolic trend suggests that an op-

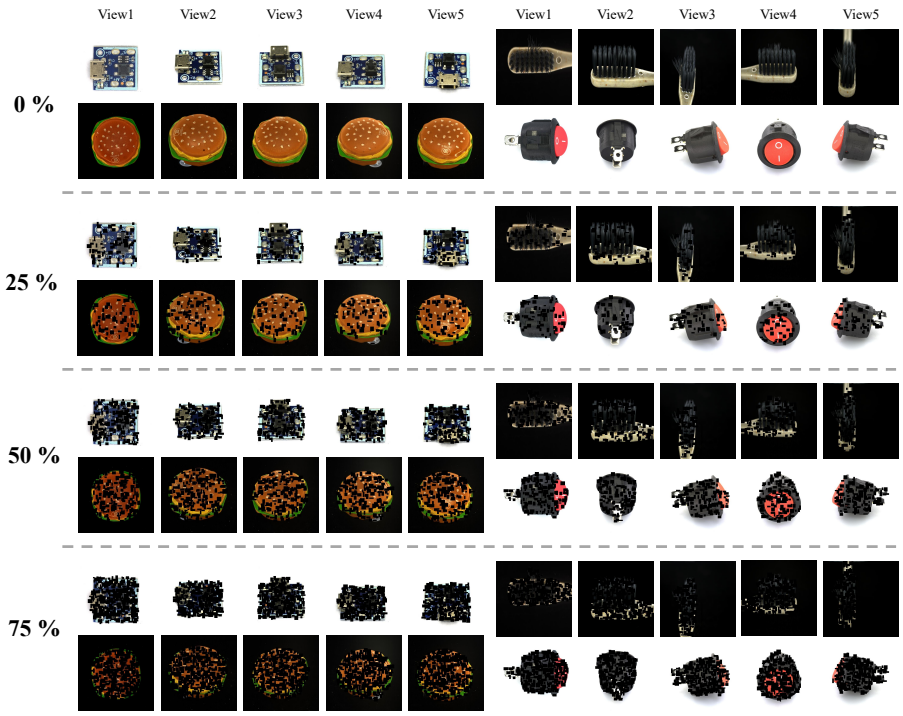

Fig. 5: Visualization of different mask ratio on RIMAD (default=50%).

timal dropout rate is essential for balancing feature robustness with information integrity, ensuring the model focuses on learning generalizable anomaly patterns.

6 Conclusion

This paper introduces IMVAD (Incomplete Multi-View Anomaly Detection), a challenging task reflecting real-world view incompleteness in industrial anomaly detection. To address this, we construct the RIMAD benchmark and propose the IMMoE framework. Our method integrates a Local Anomaly Enhancement Encoder (LAEE) to simulate anomaly patterns, a Multi-view Expert Fusion (MVEF) module to capture cross-view correlations for information reconstruction, and an Area Adaptive Loss (AAL) to prioritize challenging regions. This synergistic design maintains inter-view consistency and ensures robust detection even under severe view-incompleteness. Experimental results and heatmaps demonstrate that IMMoE achieves state-of-the-art performance on both RIMAD and Real-IAD datasets. Especially on RIMAD, we have increased the pixel-level and image-level metrics by 11.8% and 2.8%, respectively. This work underscores the necessity of multi-view synergy and provides a robust solution for comprehensive industrial inspection under incomplete information.

References

1. Aqeel, M., Sharifi, S., Cristani, M., Setti, F.: Towards real unsupervised anomaly detection via confident meta-learning. In: Proceedings of the ieee/cvf international conference on computer vision. pp. 4858–4867 (2025)
2. Cai, W., Huang, W., Cao, Y., Huang, C., Yuan, F., Zhang, B., Wen, J.: Towards vlm-based hybrid explainable prompt enhancement for zero-shot industrial anomaly detection. In: Kwok, J. (ed.) Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25. pp. 711–719. International Joint Conferences on Artificial Intelligence Organization (8 2025). https://doi.org/10.24963/ijcai.2025/80, https://doi.org/10.24963/ijcai.2025/80, main Track
3. Cheng, J., Gao, C., Zhou, J., Wen, J., Dai, T., Wang, J.: Mc3d-ad: A unified geometry-aware reconstruction model for multi-category 3d anomaly detection. In: Kwok, J. (ed.) Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25. pp. 837–845. International Joint Conferences on Artificial Intelligence Organization (8 2025). https://doi.org/10.24963/ijcai.2025/94, https://doi.org/10.24963/ijcai.2025/94, main Track
4. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023)
5. Fan, L., Fan, D., Hu, Z., Ding, Y., Di, D., Yi, K., Pagnucco, M., Song, Y.: Manta: A large-scale multi-view and visual-text anomaly detection dataset for tiny objects. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25518–25527 (2025)
6. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**(120), 1–39 (2022)
7. Gu, Z., Zhu, B., Zhu, G., Chen, Y., Ge, W., Tang, M., Wang, J.: Anomalymoe: Towards a language-free generalist model for unified visual anomaly detection. arXiv preprint arXiv:2508.06203 (2025)
8. Guo, J., Lu, S., Zhang, W., Chen, F., Li, H., Liao, H.: Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20405–20415 (2025)
9. Han, D., Pan, X., Han, Y., Song, S., Huang, G.: Flatten transformer: Vision transformer using focused linear attention. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5961–5971 (2023)
10. He, H., Bai, Y., Zhang, J., He, Q., Chen, H., Gan, Z., Wang, C., Li, X., Tian, G., Xie, L.: Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. Advances in Neural Information Processing Systems **37**, 71162–71187 (2024)
11. He, H., Zhang, J., Chen, H., Chen, X., Li, Z., Chen, X., Wang, Y., Wang, C., Xie, L.: A diffusion-based framework for multi-class anomaly detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 8472–8480 (2024)
12. He, H., Zhang, J., Tian, G., Wang, C., Xie, L.: Learning multi-view anomaly detection with efficient adaptive selection. IEEE Transactions on Multimedia (2026)
13. Hu, L., Gan, Z., Deng, L., Liang, J., Liang, L., Huang, S., Chen, T.: Replaycad: Generative diffusion replay for continual anomaly detection. In: Kwok, J. (ed.) Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25. pp. 2946–2954. International Joint Conferences on Artificial Intelligence Organization (8 2025). https://doi.org/10.24963/ijcai.2025/328, https://doi.org/10.24963/ijcai.2025/328, main Track

14. Hu, L., Huang, S.: Enhancing table structure recognition via bounding box guidance. In: International Conference on Pattern Recognition. pp. 209–225. Springer (2024)
15. Huang, C., Li, Q., Wen, J., Zhang, B.: Omni-dimensional state space model-driven sam for pixel-level anomaly detection. In: Kwok, J. (ed.) Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25. pp. 1152–1160. International Joint Conferences on Artificial Intelligence Organization (8 2025). https://doi.org/10.24963/ijcai.2025/129, https://doi.org/10.24963/ijcai.2025/129, main Track
16. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1), 79–87 (1991)
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
18. Lei, T., Chen, S., Wang, B., Jiang, Z., Zou, N.: Adapted-moe: Mixture of experts with test-time adaption for anomaly detection. In: International Conference on Intelligent Computing. pp. 427–441. Springer (2025)
19. Lin, J., Chen, Z., Liang, L., Peng, W., Huang, S.: Handwriting trajectory recovery via trajectory transformer with global radical context-aware module. In: International Conference on Pattern Recognition. pp. 182–195. Springer (2024)
20. Lin, J., Dai, G., Peng, W., Huang, S., Lao, Y., Zhang, H., Chen, T.: Atcmd-bench: Agentic traditional chinese medicine diagnosis benchmark for large language models via multi-agent simulation. Pattern Recognition **179**, 113679 (2026)
21. Liu, J., Xie, G., Wang, J., Li, S., Wang, C., Zheng, F., Jin, Y.: Deep industrial image anomaly detection: A survey. Machine Intelligence Research **21**(1), 104–135 (2024)
22. Liu, Z., Zhou, Y., Xu, Y., Wang, Z.: Simplenet: A simple network for image anomaly detection and localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20402–20411 (2023)
23. Luo, W., Cao, Y., Yao, H., Zhang, X., Lou, J., Cheng, Y., Shen, W., Yu, W.: Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9974–9983 (2025)
24. Mao, K., Lian, Y., Wang, Y., Liu, M., Zheng, N., Wei, P.: Unveiling multi-view anomaly detection: Intra-view decoupling and inter-view fusion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 12381–12389 (2025)
25. Meng, S., Meng, W., Zhou, Q., Li, S., Hou, W., He, S.: Moead: A parameter-efficient model for multi-class anomaly detection. In: European Conference on Computer Vision. pp. 345–361. Springer (2024)
26. Meng, S., Meng, W., Zhou, Q., Li, S., Hou, W., He, S.: Moead: A parameter-efficient model for multi-class anomaly detection. In: European Conference on Computer Vision. pp. 345–361. Springer (2024)
27. Miao, B., Zhang, W., Li, J., Wu, W., Tang, S., Li, Z., Shi, H., Xiao, J., Zhuang, Y.: Robust modality-incomplete anomaly detection: A modality-instructive framework with benchmark. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 7317–7326 (2025)
28. Peng, W., Huang, H., Chen, T., Ke, Q., Dai, G., Huang, S.: Globally correlation-aware hard negative generation. International Journal of Computer Vision **133**(5), 2441–2462 (2025)
29. Peng, W., Ke, Q., Liang, J., Huang, S., Chen, T.: Proxy-an loss for deep metric learning. Neural Networks p. 108254 (2025)

30. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of adam and beyond. arXiv preprint arXiv:1904.09237 (2019)
31. Riquelme, C., Puigcerver, J., Nayak, B., Mariet, Z., Pinto, M., et al.: Scaling vision with sparse mixture of experts. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 34, pp. 8583–8595 (2021)
32. Shao, Y., Wang, L., Li, C., Chen, P., Liu, Q.: Promptmoe: Generalizable zero-shot anomaly detection via visually-guided prompt mixtures. arXiv preprint arXiv:2511.18116 (2025)
33. Shazeer, N., Mirhovseini, A., Maziarz, K., Davis, A., Quoc, L., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
34. Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H.: Efficient attention: Attention with linear complexities. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3531–3539 (2021)
35. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
37. Wang, C., Zhu, W., Gao, B.B., Gan, Z., Zhang, J., Gu, Z., Qian, S., Chen, M., Ma, L.: Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22883–22892 (2024)
38. Wortsman, M., Dettmers, T., Zettlemoyer, L., Morcos, A., Farhadi, A., Schmidt, L.: Stable and low-precision training for large-scale vision-language models. Advances in Neural Information Processing Systems **36**, 10271–10298 (2023)
39. You, Z., Cui, L., Shen, Y., Yang, K., Lu, X., Zheng, Y., Le, X.: A unified model for multi-class anomaly detection. Advances in Neural Information Processing Systems **35**, 4571–4584 (2022)
40. Zavrtanik, V., Kristan, M., Skočaj, D.: Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In: ICCV. pp. 8330–8339 (2021)
41. Zhang, J., Chen, X., Wang, Y., Wang, C., Liu, Y., Li, X., Yang, M.H., Tao, D.: Exploring plain vit reconstruction for multi-class unsupervised anomaly detection. arXiv preprint arXiv:2312.07495 (2023)
42. Zhang, X., Li, S., Li, X., Huang, P., Shan, J., Chen, T.: 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 (2023)
43. Zhou, C., Paffenroth, R.C.: Anomaly detection with robust deep autoencoders. In: SIGKDD. pp. 665–674 (2017)