

Multi-view Patch Inference with Frozen DINOv3 for Industrial Anomaly Detection

Chaoqun Wang

Beijing Union University

Wenjing Zhang

University of Chinese Academy of Sciences

Bo Qi

Beijing Union University

Xiaoyu Huang

Beijing Union University

Ning He

xxthening@bnu.edu.cn

Beijing Union University

Research Article

Keywords: Industrial Anomaly Detection, Vision Transformers, Feature Reconstruction, Self-supervised Learning

Posted Date: July 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10224105/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Multimedia Systems on October 10th, 2026. See the published version at <https://doi.org/10.1007/s00530-026-02685-0>.

Multi-view Patch Inference with Frozen DINOv3 for Industrial Anomaly Detection

Chaoqun Wang¹, Wenjing Zhang³, Bo Qi², Xiaoyu Huang¹,
Ning He^{1*}

^{1*}College of Smart City, Beijing Union University, Beijing, China.

²College of Robotics, Beijing Union University, Beijing, China.

³School of Engineering Science, University of Chinese Academy of
Sciences, Beijing, 100049, China.

*Corresponding author(s). E-mail(s): xxthening@buu.edu.cn;

Contributing authors: 20241081210203@buu.edu.cn;

zhangwenjing242@mails.ucas.ac.cn; 20241083510910@buu.edu.cn;

20242085410205@buu.edu.cn;

Abstract

Industrial anomaly detection is a critical task in automated visual inspection systems, aiming to identify subtle defects without relying on large-scale anomaly annotations. Recently, pretrained vision transformers such as DINO have demonstrated strong feature representation capabilities and have been widely adopted in this domain. However, most existing methods rely on single-grid patch partitioning, which suffers from boundary information loss and limited local context modeling, resulting in degraded performance on small or boundary-located defects. In this paper, we propose a simple yet effective framework based on a frozen DINOv3 backbone for industrial anomaly detection. To address the limitations of conventional patch-based inference, we introduce a Multi-view Patch Inference (MVPI) strategy, which constructs multiple overlapping spatial views of an input image. Each view is independently processed by DINOv3, and overlapping regions are aggregated to enhance feature consistency and boundary robustness. On top of the extracted multi-view features, we further propose a Multi-scale Reconstruction Fusion Transformer (MRFT) to model normal feature distributions via cosine similarity-based reconstruction learning. The anomaly score is computed based on reconstruction residuals, enabling effective discrimination between normal and defective regions. Extensive experiments on three benchmark datasets, including MVTec AD, BTAD, and the Ultrachip dataset, demonstrate that the proposed method consistently outperforms strong baselines

in both image-level and pixel-level anomaly detection tasks. The results validate the effectiveness of combining multi-view inference with frozen vision transformer representations for robust industrial anomaly detection. The code will be released at <https://github.com/chaoqunWang666>.

Keywords: Industrial Anomaly Detection, Vision Transformers, Feature Reconstruction, Self-supervised Learning

1 Introduction

Industrial Anomaly Detection (IAD) is a fundamental task in automated visual inspection systems, aiming to identify defects from images without requiring large-scale anomaly annotations. In real industrial scenarios, such as ultrasonic chip inspection, defects are typically small-scale, irregular, and embedded in highly noisy grayscale backgrounds. Moreover, spatial misalignment and appearance variability further increase the difficulty of robust defect localization. These challenges require models to simultaneously achieve strong representation capability, spatial robustness, and computational efficiency.

Recent advances in IAD are largely driven by pretrained vision foundation models. Representative paradigms include memory bank-based methods [1–4], feature distillation approaches [5], and reconstruction-based methods [6]. Although these methods achieve promising results, they still suffer from several limitations. Memory-based methods introduce high storage and retrieval costs. Distillation-based approaches depend heavily on teacher-student alignment quality. Reconstruction-based frameworks often rely on complex architectural designs such as masked image modeling or heavy encoder-decoder structures, which significantly increase computational overhead and may degrade local structural fidelity.

More recently, transformer-based pretrained models, particularly DINO-style vision transformers, have demonstrated strong potential for industrial anomaly detection due to their powerful semantic representation capability [7]. However, existing approaches typically rely on single-grid patch partitioning during inference, which leads to boundary information loss and insufficient local context modeling. As a result, subtle defects located at patch boundaries are often missed or weakened in the final anomaly response.

To address the above limitations, we revisit the design of patch-based inference and feature reconstruction in a more lightweight and efficient manner. Instead of introducing complex masking strategies or heavy decoder structures, we argue that robust anomaly detection can be achieved through better spatial sampling and hierarchical feature refinement on frozen vision representations.

In this work, we propose a simple yet effective framework built upon a frozen DINOv3 backbone [8]. To enhance spatial robustness, we introduce a Multi-view Patch Inference (MVPI) strategy, which generates multiple overlapping spatial views of an input image. Each view is independently processed by the DINOv3 backbone, and overlapping regions are aggregated to improve boundary consistency and reduce spatial

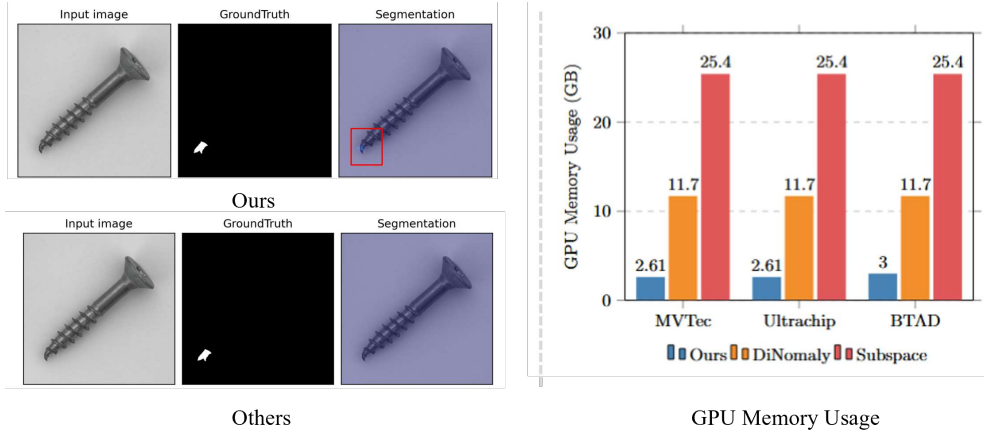


Fig. 1 Motivation of the proposed framework. **Left panel:** Visual comparison of anomaly localization on a subtle defect. Conventional single-view inference (bottom) fails to capture the tiny defect due to boundary information loss, whereas our Multi-view Patch Inference (MVPI) strategy (top) achieves accurate localization. **Right panel:** Comparison of GPU memory usage (GB) across MVTEC, Ultrachip, and BTAD datasets. Benefiting from our lightweight Multi-scale Reconstruction Fusion Transformer (MRFT), our method drastically reduces computational overhead compared to DiNomaly and Subspace, ensuring high deployment efficiency.

aliasing effects. As illustrated in the left panel of Fig. 1, by mitigating boundary information loss, our MVPI strategy successfully localizes subtle and hard defects (e.g., the tiny scratch on a screw tip) that are completely missed by conventional single-view methods.

Furthermore, we design a Multi-scale Reconstruction Fusion Transformer (MRFT), which performs hierarchical feature refinement across multiple transformer stages. Specifically, shallow, intermediate, and deep reconstruction layers are used to capture texture-level, structure-level, and semantic-level information, respectively. The multi-scale outputs are then fused to produce a stable reconstruction of normal feature representations. Crucially, unlike existing paradigms that incur heavy computational burdens, our lightweight multi-scale transformer architecture significantly reduces computational overhead. As demonstrated in the right panel of Fig. 1, our framework drastically lowers GPU memory usage across multiple benchmarks (MVTEC, Ultrachip, and BTAD) compared to state-of-the-art counterparts, achieving an optimal trade-off between detection accuracy and deployment efficiency.

The main contributions of this work are summarized as follows:

- **Multi-view Patch Inference Strategy:** We propose a Multi-view Patch Inference (MVPI) mechanism that constructs overlapping spatial views for each input image. This design mitigates boundary information loss in conventional patch-based inference and improves robustness for small and boundary-located defects.

- **Multi-scale Reconstruction Fusion Transformer:** We introduce a lightweight MRFT module that performs hierarchical feature reconstruction across multiple transformer layers. By integrating shallow-to-deep representations, the model achieves more stable and discriminative normal feature modeling.
- **Efficient DINOv3-based Anomaly Detection Framework:** We build a simple yet powerful anomaly detection framework based on a frozen DINOv3 backbone. The proposed method achieves strong performance on MVTec AD, BTAD, and the challenging Ultrachip dataset, demonstrating the effectiveness of combining multi-view inference with multi-scale feature reconstruction.

2 Related Work

2.1 Large Language and Vision-Language Models for Anomaly Detection

The emergence of large language models (LLMs) and vision-language models (VLMs) has opened a new paradigm for zero-shot and few-shot industrial anomaly detection. Models such as CLIP[9], GPT-4V[10], and LLaVA leverage massive image-text pre-training to align visual and linguistic representations, enabling anomaly detection through natural language prompts without any task-specific training data. In the zero-shot setting, these models directly compare test images against textual descriptions of "normal" or "anomalous" states via prompt engineering, bypassing the need for normal reference samples entirely[11]. For few-shot scenarios, in-context learning capabilities allow LLMs to reason about anomalies using only a handful of annotated examples as contextual cues. Recent works further enhance this paradigm by integrating domain-specific knowledge bases, chain-of-thought reasoning, or adapter tuning to bridge the semantic gap between open-world pretraining and fine-grained industrial defect patterns. Despite their remarkable flexibility, VLM-based methods still face challenges in pixel-level localization precision and computational efficiency, often requiring complementary mechanisms to meet real-time industrial deployment requirements.

2.2 Unsupervised Anomaly Detection

Unsupervised anomaly detection methods operate under the assumption that only normal samples are available during training. This paradigm is broadly categorized into unified multi-class approaches and single-class approaches, each addressing distinct deployment scenarios.

2.2.1 Unified Multi-Class Unsupervised Methods

Unified multi-class methods aim to train a single model capable of detecting anomalies across multiple product categories or domains simultaneously, eliminating the need for category-specific retraining. Representative works leverage shared feature spaces, prompt tuning, or adapter modules to align diverse normal distributions into a common reference frame[12–17]. These approaches significantly reduce deployment overhead in multi-product manufacturing lines but often face trade-offs between

cross-category generalization and per-category sensitivity, particularly when inter-class variance is high.

2.2.2 Single-Class Unsupervised Methods

Single-class methods remain the dominant paradigm in IAD due to their superior performance on individual product lines. They are further divided into four sub-categories:

Teacher-Student Architectures.

Teacher-Student (T-S) frameworks employ a pre-trained teacher network to extract target features from normal samples, while a student network is trained to mimic these representations exclusively on normal data [5]. Anomalies are localized via feature discrepancy between the two networks at inference time. Although T-S models avoid explicit memory banks, they suffer from capacity mismatch challenges: excessive student capacity leads to over-generalization on anomalous regions, diminishing anomaly score distinctiveness, while insufficient capacity hinders faithful feature alignment.

Memory Bank-based Methods.

These methods construct reference profiles from normal image embeddings to enable non-parametric anomaly scoring. PaDiM [18] models multivariate Gaussian distributions of patch features, while PatchCore [19] builds a coreset of normal features for nearest-neighbor retrieval. Despite exceptional benchmark performance, memory bank approaches incur substantial GPU memory costs for storing large feature banks and suffer from high inference latency due to pixel-wise neighbor search. Statistical variants also rely on ideal distribution assumptions that may fail in noisy industrial environments.

One-Class Classification (OCC).

OCC methods learn a compact boundary or hypersphere enclosing normal feature representations in latent space[20]. Classical approaches like Deep SVDD minimize the volume of a bounding hypersphere around normal embeddings, treating points outside as anomalies. Recent extensions integrate deep feature extractors and adaptive margin mechanisms to handle complex manifolds. However, OCC methods often struggle with high-dimensional feature spaces where the curse of dimensionality causes distance metrics to lose discriminative power, and they typically lack fine-grained localization capability without additional architectural modifications.

Reconstruction-based Methods.

Reconstruction methods train autoencoders or GANs [21] exclusively on normal samples [22], identifying anomalies via input-reconstruction discrepancies. Early approaches suffer from the "identity mapping" problem, where networks generalize too well to anomalous inputs. Advanced methods like DRAEM [23], CutPaste [24], and RealNet [25] mitigate this through pseudo-anomaly generation or diffusion-synthesized anomalies, but introduce cumbersome augmentation pipelines and computational

overhead. In contrast, our approach adheres to a minimalist pure reconstruction philosophy: operating entirely in semantic feature space and explicitly minimizing cosine distance on a unit hypersphere, we bypass identity mapping without relying on complex pseudo-anomaly generation.

3 Method

3.1 Overview

We propose a Multi-view Patch Inference framework with a Multi-scale Reconstruction Fusion Transformer (MRFT) for industrial anomaly detection, as illustrated in Fig. 2. The framework consists of three components: (1) a frozen DINOv3 backbone for hierarchical feature extraction, (2) a Multi-view Patch Inference (MVPI) strategy for robust spatial representation, and (3) a multi-scale reconstruction fusion module for anomaly modeling. Given an input image x , the model first extracts multi-layer

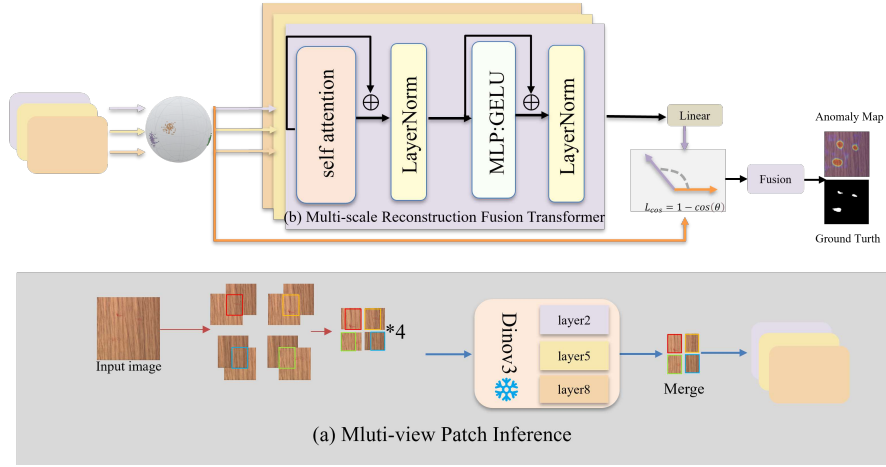


Fig. 2 The overall architecture of our Multi-view Patch Inference with a Multi-scale Reconstruction Fusion Transformer Framework

features using a pretrained DINOv3 backbone. Then, multi-view patch inference is applied to enhance spatial robustness. Finally, multi-scale transformer-based reconstruction is performed to estimate normal feature distributions and compute anomaly scores.

3.2 DINOv3 Foundation Model with Enhanced Token Semantics

We adopt a pretrained DINOv3 vision transformer as a frozen feature extractor. Compared with earlier DINO versions, DINOv3 introduces several key architectural improvements that are particularly beneficial for industrial anomaly detection.

3.2.1 Register Tokens for Global Semantic Anchoring

DINOv3 introduces additional register tokens, which act as global semantic anchors during self-attention computation. Unlike standard patch tokens, register tokens do not correspond to specific spatial locations but instead aggregate global context information. Formally, given input tokens $\mathbf{T} \in \mathbb{R}^{N \times d}$, register tokens $\mathbf{R} \in \mathbb{R}^{k \times d}$ are appended:

$$\mathbf{T}' = [\mathbf{T}; \mathbf{R}] \quad (1)$$

Self-attention is then computed over the extended token set:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) \mathbf{V} \quad (2)$$

where register tokens improve global dependency modeling and reduce spatial ambiguity in defect-prone regions.

3.2.2 High-Capacity Backbone (7B-scale Representation)

DINOv3 benefits from a significantly larger parameter scale compared to previous versions, reaching up to billions of parameters, enabling stronger representation of fine-grained textures and structural variations. This large-scale pretraining improves texture sensitivity, structural consistency, and cross-category generalization, which are critical for industrial anomaly detection under domain shift.

3.2.3 Multi-layer Feature Selection Strategy

We extract intermediate features from layers $l \in \{2, 5, 8\}$:

- **Layer 2 (Low-level features):** Captures edge structures and fine texture gradients; sensitive to micro-defects.
- **Layer 5 (Mid-level features):** Encodes part-level structure and local geometric consistency; handles defect boundary transitions.
- **Layer 8 (High-level features):** Represents semantic consistency and object-level structure; enables global anomaly suppression.

The fusion of these layers provides a complementary representation spanning from local texture to global semantics.

3.3 Multi-view Patch Inference (MVPI)

To address boundary information loss in conventional patch-based inference, we propose a Multi-view Patch Inference strategy implemented at the feature extraction stage, as illustrated in Fig.2 (a).

3.3.1 Spatial Multi-view Construction

Given an input feature map $\mathbf{X}$, we construct a set of overlapping views $\mathcal{V} = \{x_1, x_2, \dots, x_N\}$, where each view is generated via a sliding-window mechanism with overlap constraint:

$$x_i = \mathcal{C}(\mathbf{X}; s_i, s_i + p) \quad (3)$$

Here, p denotes the patch size, s_i represents the spatial offset with overlap, and the overlap ratio controls the redundancy level.

3.3.2 Multi-view Feature Encoding

Each view is independently processed by the frozen DINOv3 backbone:

$$\mathbf{F}_i^l = \phi_l(x_i), \quad l \in \{2, 5, 8\} \quad (4)$$

This results in redundant feature observations of the same spatial region under different receptive contexts.

3.3.3 Cross-view Feature Aggregation

For overlapping regions, features are aggregated using weighted averaging:

$$\mathbf{F} = \sum_{i=1}^N w_i \mathbf{F}_i \quad (5)$$

where weights w_i are implicitly determined by spatial coverage frequency. This mechanism improves boundary consistency, robustness to spatial misalignment, and sensitivity to small defects.

3.3.4 Effect of MVPI

MVPI effectively transforms single-view inference into an implicit ensemble of multiple spatial views, reducing the risk of missing anomalies located at patch boundaries.

3.4 Multi-scale Reconstruction Fusion Transformer (MRFT)

3.4.1 Motivation

Single-scale reconstruction methods are limited in capturing hierarchical feature dependencies. To address this, we propose a multi-scale transformer-based reconstruction strategy, as illustrated in Fig.2 (b).

3.4.2 Hierarchical Reconstruction Formulation

Given input token features $\mathbf{z}$, we define a three-stage transformation:

$$\mathbf{z}_1 = f_1(\mathbf{z}), \quad \mathbf{z}_2 = f_2(\mathbf{z}_1), \quad \mathbf{z}_3 = f_3(\mathbf{z}_2) \quad (6)$$

where each f_i is a lightweight self-attention encoder layer:

$$f_i(\mathbf{x}) = \text{LN}(\mathbf{x} + \text{MSA}(\mathbf{x})) + \text{FFN}(\mathbf{x}) \quad (7)$$

where LN denotes Layer Normalization, MSA denotes Multi-head Self-Attention, and FFN denotes Feed-Forward Network.

3.4.3 Multi-scale Feature Fusion

The final reconstruction output is obtained via fusion:

$$\hat{\mathbf{z}} = \sum_{i=1}^3 \alpha_i \mathbf{z}_i, \quad \sum_{i=1}^3 \alpha_i = 1 \quad (8)$$

This design enables low-level refinement (texture), mid-level correction (structure), and high-level stabilization (semantics).

3.4.4 Jitter Regularization

To improve robustness, we introduce feature jitter:

$$\mathbf{z}' = \mathbf{z} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \quad (9)$$

which prevents overfitting to exact normal feature patterns.

3.5 Anomaly Score Computation

The anomaly score is defined as the reconstruction deviation:

$$S = 1 - \cos(\mathbf{z}, \hat{\mathbf{z}}) \quad (10)$$

Pixel-level anomaly maps are obtained by reshaping token-wise scores and interpolating to image resolution. The image-level anomaly score is computed via top- k aggregation:

$$S_{\text{img}} = \frac{1}{K} \sum_{i \in \text{TopK}} S_i \quad (11)$$

Gaussian smoothing is subsequently applied to improve spatial continuity.

4 Experiments

4.1 Experimental Setup

Datasets. To comprehensively evaluate the robustness and generalizability of our proposed method, we conduct extensive experiments on three distinct industrial anomaly detection datasets. 1) **MVTec AD**[26]: A standard benchmark containing 15 categories of industrial products, widely used to evaluate basic detection performance. 2) **BTAD**[27]: Comprises 3 categories of real-world industrial scenarios, heavily utilized to assess the model’s environmental robustness. 3) **UltraChip**[28]: A highly challenging dataset focusing on micro-defect detection in semiconductor chips, which we employ to validate our method’s generalization capability on extremely fine-grained microscopic anomalies.

Evaluation Metrics. Following standard anomaly detection protocols, we employ three widely-accepted metrics to quantitatively evaluate our model. We use Image-level Area Under the Receiver Operating Characteristic Curve (I-AUC) for anomaly classification, Pixel-level Area Under the ROC (P-AUC) to assess anomaly localization precision, and Per-Region Overlap (PRO) to measure the region-wise overlap rate of the predicted anomalous areas.

Implementation Details. Most of our experiments are conducted on a single NVIDIA V100 GPU (24GB). All input images are uniformly resized to 448×448 . To construct a multi-scale receptive field, we extract intermediate features from the 2nd, 5th, and 8th layers of the encoder. During the inference stage, we perform a topological smoothing operation using a Gaussian filter with a constant standard deviation ($\sigma = 1.5$) to suppress background noise, and establish robust image-level anomaly scores by calculating the mean of the Top- K ($K = 100$) extreme values.

4.2 Comparison with State-of-the-Art Methods

To comprehensively evaluate our framework, we integrate quantitative metrics and qualitative visualizations. We first benchmark our method on the highly challenging real-world dataset, and subsequently verify its generalizability on standard IAD benchmarks.

Performance on Real-World UltraChip Dataset. We prioritize the evaluation on the UltraChip dataset, which poses extreme challenges due to severe structural noise, limited labeled data, and highly diverse, visually imperceptible microscopic defects. As summarized in Table 1, while traditional methods suffer sharp performance degradation from heavy background interference, our method successfully maintains a state-of-the-art mean Image-AUC of 81.9%. Notably, our approach outperforms recent models like URD and SuperSimpleNet by significant margins. Furthermore, as demonstrated perfectly in the qualitative results in Figure 3, our framework precisely highlights actual defect areas without being inundated by widespread false-positive background noise (avoiding massive ”red” regions), proving its exceptional capability to extract intrinsic normal patterns in non-ideal industrial environments.

Generalizability on Standard Benchmarks (MVTec AD & BTAD). To prove that our method does not overfit to specific noisy datasets, we benchmark it on

Table 1 Detailed quantitative comparison on the UltraChip dataset. We report the performance across all three specific chip categories.

Class	Four-Lead			Bonding			Two-Lead			Mean		
Method	I-AUC	P-AUC	PRO	I-AUC	P-AUC	PRO	I-AUC	P-AUC	PRO	I-AUC	P-AUC	PRO
REB[29]	64.0	83.2	33.9	83.2	0.285	43.6	66.1	87.3	33.2	71.1	68.3	36.9
FUNAD[30]	72.5	82.5	39.0	85.1	94.8	44.5	76.4	81.4	29.5	78.0	86.2	37.6
RAS[31]	77.9	87.7	36.0	64.9	80.6	44.4	77.7	76.4	29.2	73.5	81.6	36.5
MSFR[22]	36.7	55.4	19.6	53.6	57.4	16.4	50.1	51.5	18.5	46.8	54.8	18.1
FOD[32]	79.1	80.2	32.6	79.9	89.0	54.7	79.8	74.9	18.6	79.6	81.3	35.3
RD++[5]	71.9	84.7	50.8	64.2	86.1	41.4	77.7	83.9	44.7	71.2	84.9	45.6
HGAD[33]	74.3	83.5	46.6	78.1	88.8	43.1	75.8	83.2	27.5	76.0	85.1	39.1
SuperSimpleNet[34]	72.9	67.1	40.3	78.0	81.5	56.2	79.4	68.7	33.2	76.8	72.5	43.2
URD[35]	66.7	82.6	29.4	86.8	95.5	62.8	68.8	79.2	21.5	74.1	85.7	37.9
DiNomaly[12]	83.9	83.9	47.6	64.9	87.6	48.1	80.3	84.6	44.4	76.4	85.4	46.7
Subspace[1]	73.5	85.5	34.1	27.6	73.1	42.9	55.7	85.3	29.6	52.3	81.3	35.5
Ours	80.1	84.3	45.7	82.5	92.3	58.9	83.3	82.2	34.6	81.9	86.3	46.4

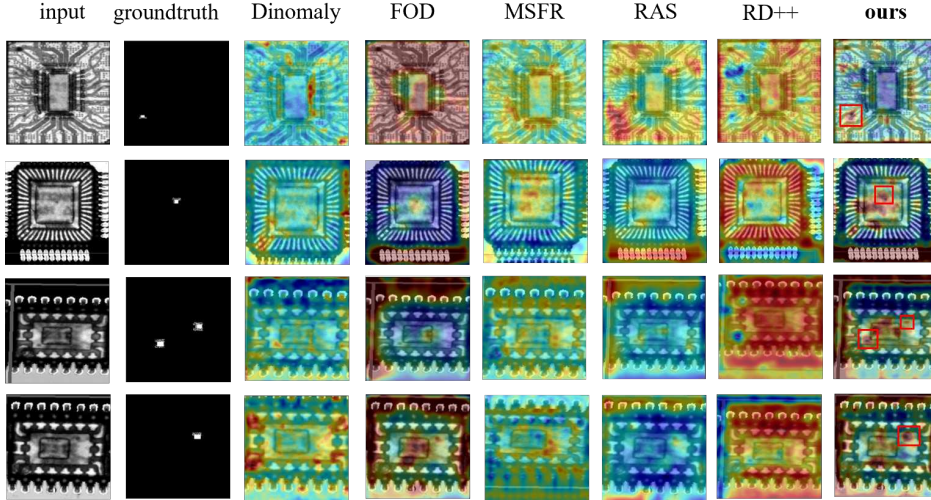


Fig. 3 Qualitative comparison of anomaly localization on the challenging UltraChip dataset.

the standard MVTec AD and BTAD datasets. As presented in Table 2 3 , despite its architectural minimalism, our method achieves the most comprehensive performance on MVTec AD, securing the highest scores across all two metrics (99.2% P-AUC, and 96.0% PRO). This represents a clear improvement over SOTA methods like Dinomaly and Subspace. To intuitively demonstrate this, Figure 4 shows that our generated heatmaps align tightly with ground-truth masks with almost zero background

noise. Moreover, Table 4 confirms our robustness against varying lighting and complex topologies typical of factory settings in the BTAD dataset, where we maintain a consistent advantage in average metrics (97.3% I-AUC).

Table 2 Quantitative comparison on the MVTec AD dataset.

Method	I-AUC	P-AUC	PRO
REB[29]	99.1	98.4	92.5
URD[35]	98.7	98.8	96.0
SuperSimpleNet[34]	97.4	98.5	91.0
FUNAD[30]	99.2	98.1	94.2
RD++[5]	98.2	99.2	94.9
FOD[32]	98.9	98.3	93.5
HGAD[33]	98.4	97.9	94.2
RAS[31]	98.5	97.5	93.9
MSFR[22]	98.4	97.3	93.0
Dinomaly [12]	99.6	98.3	94.8
Subspace [1]	99.1	98.1	94.9
Ours	99.5	99.2	96.0

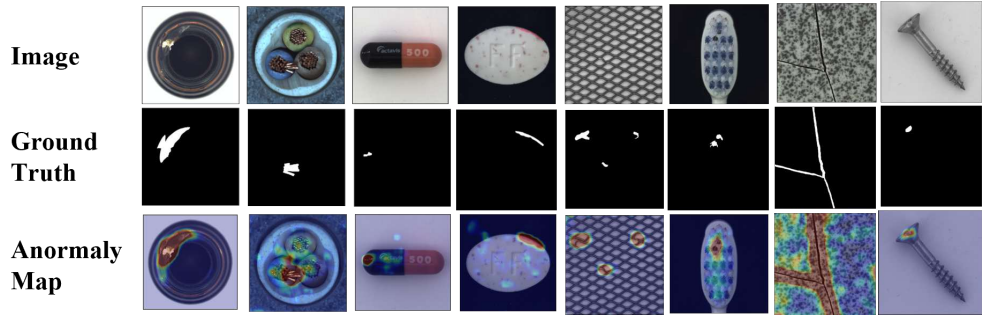


Fig. 4 Qualitative anomaly localization results of our proposed method on the MVTec AD dataset. For each category, we present the raw input image, the ground-truth anomaly mask, and the highly precise anomaly heatmap generated by our framework.

4.3 Ablation Studies

To rigorously validate the efficacy of each proposed module and verify the rationality of our structural design, we conduct comprehensive ablation studies on the MVTec AD datasets.

As demonstrated in Table 5, we conducted an ablation study to evaluate the impact of different feature layer combinations from DINOv3 on anomaly detection performance for the *screw* category. The configuration utilizing layers {2, 5, 8} achieves the highest

Table 3 MVTec Performance Metrics (AUC) — DINOv3 vs. DINOv3+MVPI Results

Category	DINOv3			DINOv3+MVPI		
	I-AUC	P-AUC	PRO	I-AUC	P-AUC	PRO
hazelnut	100.0	99.5	96.7	100.0	99.5	96.9
metal_nut	100.0	99.4	97.1	100.0	99.3	97.0
bottle	100.0	99.2	94.0	100.0	99.2	97.0
cable	99.7	99.2	99.0	99.7	99.1	99.5
capsule	98.4	99.3	89.0	98.7	99.3	89.2
carpet	100.0	99.7	98.3	100.0	99.7	98.0
grid	100.0	99.1	96.3	100.0	99.2	96.0
leather	100.0	99.4	98.4	100.0	99.5	98.5
pill	98.3	98.7	96.9	98.6	99.0	97.1
screw	95.1	99.3	95.8	98.6	99.5	96.5
tile	100.0	98.8	95.5	100.0	98.8	95.7
toothbrush	100.0	99.2	89.8	100.0	99.3	92.5
transistor	99.7	98.7	96.3	99.7	98.5	96.3
wood	97.5	97.0	92.1	97.5	97.0	92.3
zipper	99.9	98.9	96.4	99.9	99.3	97.6
Mean	99.2	99.0	95.4	99.5	99.2	96.0

Table 4 Quantitative comparison on the BTAD dataset per class.

Class	Class 01			Class 02			Class 03			Average		
Method	I-AUC	P-AUC	PRO	I-AUC	P-AUC	PRO	I-AUC	P-AUC	PRO	I-AUC	P-AUC	PRO
FOD[32]	95.8	97.0	62.2	96.2	97.8	65.4	96.0	97.7	99.0	96.0	97.5	75.5
Subspace[1]	100	98.9	84.7	91.2	96.8	80.4	99.6	98.5	56.1	96.9	98.0	73.7
Dinomaly[12]	96.4	97.2	74.0	89.1	96.5	56.6	99.8	99.8	99.3	95.1	97.9	76.6
REB[29]	99.6	94.7	70.0	88.5	95.6	66.4	99.8	99.7	80.6	96.0	97.2	72.4
HGAD[33]	92.6	94.9	70.5	93.3	96.7	59.3	97.0	99.1	97.9	94.1	96.9	75.9
RAS[31]	92.7	96.9	70.3	92.3	94.7	55.5	99.1	99.0	99.0	94.7	97.0	74.9
Ours	96.5	96.8	63.6	95.5	97.8	67.5	99.9	99.9	99.0	97.3	98.2	76.7

Table 5 Ablation study on feature layer selection for the *screw* category.

Feature Layers	I-AUC	P-AUC	PRO
{3, 6, 9}	93.9	99.3	95.6
{4, 7, 10}	92.7	99.2	95.7
{2, 5, 8}	95.2	99.2	95.8

I-AUC (95.2%) and PRO (95.8%), while maintaining a competitive P-AUC (99.2%). Although the $\{3, 6, 9\}$ setting yields a marginally superior P-AUC (99.3%), it suffers from a noticeable degradation in image-level discrimination (I-AUC drops by 1.3%). We attribute the superior overall performance of layers $\{2, 5, 8\}$ to their balanced representation: this specific combination effectively captures fine-grained shallow texture and structural details from the early layers, while simultaneously preserving rich high-level semantic information from the deeper layers. Consequently, we adopt $\{2, 5, 8\}$ as the optimal feature layer configuration for DINOv3 in our final model.

Rationale for Multi-Scale Layer Selection. Given the diverse manifestations of defects—from tiny cracks to large structural missing parts—single-scale features are highly inadequate. We meticulously select features from the 2nd, 5th, and 8th layers of the frozen DINOv3 backbone. As demonstrated in Table 6, this $\{2, 5, 8\}$ configuration yields the best results (99.5% I-AUC unconditionally). The underlying rationale is complementary: Layer 2 is shallow enough to capture fine-grained textural anomalies and high-frequency noise typical in UltraChip; Layer 5 acts as an intermediary bridge for semantic transitions; Layer 8 is deep enough to capture local structural semantics without becoming too abstract or losing spatial boundary resolution. In contrast, using only deeper layers like $\{5, 8\}$ drops the average I-AUC to 98.30% due to the loss of fine-grained details, ensuring holistic anomaly localization.

Table 6 Performance comparison of multi-scale layer selections on the MVTec AD dataset.

Selected Layers	Category	I-AUC	P-AUC	PRO
$\{5, 8\}$	Average	98.30	97.60	95.00
	Cable	94.79	98.78	97.40
$\{2, 5, 8\}$	Average	99.20	99.60	97.40
	Cable	99.50	99.30	99.00

Superiority of Top- K Score Aggregation. Finally, we evaluate strategies for mapping pixel-level anomaly maps to global image-level scores. In real-world industrial scenarios, background reflections and sensor artifacts often induce isolated high-frequency noise spikes. As demonstrated in Table 7, naive Global Max-pooling is highly vulnerable to these single-pixel outliers, severely inflating the false positive rate. Conversely, Global Average-pooling over-smooths the score map, drowning out subtle microscopic defects within dominant normal regions.

It is worth noting that before aggregation, our pipeline applies Gaussian smoothing ($\sigma = 1.5$) and squared enhancement to suppress high-frequency jitter and amplify salient defect responses. Building on this refined score map, our proposed Top- K Pooling ($K = 100$) leverages a rank-based selection mechanism that inherently excludes isolated noise spikes while preserving the spatial density signals from actual defect regions. This robust aggregation strategy achieves the best trade-off between noise suppression and defect sensitivity, maximizing overall detection performance.

Table 7 Comparison of image-level score aggregation strategies on MVTec AD. All methods share identical pixel-level preprocessing (Gaussian blur + squared enhancement) before aggregation.

Aggregation Strategy	I-AUC	P-AUC	PRO
Global Max-pooling	98.9	98.8	96.2
Global Average-pooling	98.3	99.2	95.0
Top-K Pooling ($K = 100$)	99.5	99.1	95.8

5 Conclusion

In this paper, we propose a simple yet effective Multi-view Patch Inference with Reconstruction Fusion Transformer (MVPI-RFT) framework for industrial anomaly detection. By introducing a Multi-view Patch Inference (MVPI) strategy and a Multi-scale Reconstruction Fusion Transformer (MRFT) mechanism, our approach effectively addresses the limitations of conventional single-grid patch partitioning while completely eliminating the need for cumbersome memory banks and complex distillation-based architectures. Extensive experiments on standard benchmarks, including MVTec AD, BTAD, and the highly challenging UltraChip dataset, demonstrate that our method achieves state-of-the-art performance in both image-level and pixel-level anomaly detection, with strong robustness and precise localization ability, while maintaining a lightweight and memory-efficient inference pipeline.

Despite these advantages, the use of a frozen DINOv3 backbone introduces additional computational overhead, leading to relatively higher inference latency. Future work will focus on designing more efficient backbone-adaptive strategies to reduce computational cost and improve real-time applicability. Interestingly, our method also exhibits promising zero-shot generalization behavior on the UltraChip dataset, suggesting that multi-view feature aggregation with frozen vision transformers may implicitly capture transferable normality priors. This observation opens up an interesting direction for extending the framework toward zero-shot or few-shot anomaly detection scenarios.

Lastly, our current framework still operates under a conventional single-class training paradigm, which requires training separate models for different product categories. In future work, we aim to overcome this limitation by developing a unified multi-class anomaly detection framework. We further plan to explore more compact and decoupled feature representation strategies to enhance generalization across diverse industrial categories, while also improving inference efficiency for deployment on edge devices.

Declarations

Availability of data and material

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors did not receive support from any organization for the submitted work.

Authors' contributions

Chaoqun Wang: Conceptualization, Methodology, Formal Analysis, Writing – Original Draft, Project Administration. Wenjing Zhang: Validation, Investigation, Data Curation, Writing – Review & Editing. Bo Qi: Resources, Visualization. Xiaoyu Huang: Resources, Visualization. Ning He: Supervision, Funding Acquisition, Final Approval. All authors read and approved the final manuscript.

5.1 Acknowledgments

This work was supported by the Key Research and Development Project of the Ministry of Science and Technology: Ultrasonic Microscope (2023YFF0716500)

References

- [1] Lendering, C., Akdag, E., Bondarev, E.: Subspacead: Training-free few-shot anomaly detection via subspace modeling. arXiv preprint arXiv:2602.23013 (2026)
- [2] Lin, Y., Yan, H., Tong, X., Chang, Y., Wang, H., Zhou, Z., Gao, S., Wang, Y., Zhang, W.: Commonality in few: Few-shot multimodal anomaly detection via hypergraph-enhanced memory. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 7015–7023 (2026)
- [3] Wang, Q., Guo, J., Gao, S., Wang, H., Xiong, L., Hu, J., Guo, H., Zhang, W.: Search is all you need for few-shot anomaly detection. arXiv preprint arXiv:2504.11895 (2025)
- [4] Xu, C., Lv, C., Chen, Q., Zhang, F., Zhang, Z.: Mrad: Zero-shot anomaly detection with memory-driven retrieval. arXiv preprint arXiv:2602.00522 (2026)
- [5] Tien, T.D., Nguyen, A.T., Tran, N.H., Huy, T.D., Duong, S., Nguyen, C.D.T., Truong, S.Q.: Revisiting reverse distillation for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24511–24520 (2023)
- [6] Deng, H., Li, X.: 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 (2022)

- [7] Hou, Y., Li, P., Liu, Z., Wang, Y., Ruan, Y., Qiu, J., Xu, K.: Visualad: Language-free zero-shot anomaly detection via vision transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21346–21356 (2026)
- [8] Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025). <https://arxiv.org/abs/2508.10104>
- [9] Zhou, Q., Pang, G., Tian, Y., He, S., Chen, J.: Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In: International Conference on Learning Representations, vol. 2024, pp. 49705–49737 (2024)
- [10] Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M., Wang, J.: Anomalygpt: Detecting industrial anomalies using large vision-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 1932–1940 (2024)
- [11] Fučka, M., Zavrtanik, V., Skočaj, D.: Anomalyvfm—transforming vision foundation models into zero-shot anomaly detectors. arXiv preprint arXiv:2601.20524 (2026)
- [12] 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)
- [13] 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)
- [14] Lee, Y., Lim, H., Jang, S., Yoon, H.: Uniformly: Towards task-agnostic unified framework for visual anomaly detection. *Pattern Recognition* **169**, 111820 (2026)
- [15] Lu, R., Wu, Y., Tian, L., Wang, D., Chen, B., Liu, X., Hu, R.: Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems* **36**, 8487–8500 (2023)
- [16] Roming, L., Lehnerer, F., Funk, J.V., Michel, A., Maier, G., Längle, T., Beyerer, J.: Superadd: Training-free class-agnostic anomaly segmentation—cvpr 2026 vand 4.0 workshop challenge industrial track. arXiv preprint arXiv:2605.14808 (2026)
- [17] 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)

- [18] Defard, T., Setkov, A., Loesch, A., Audigier, R.: Padim: a patch distribution modeling framework for anomaly detection and localization. In: International Conference on Pattern Recognition, pp. 475–489 (2021). Springer
- [19] Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14318–14328 (2022)
- [20] 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)
- [21] Fang, Q., Su, Q., Lv, W., Xu, W., Yu, J.: Boosting fine-grained visual anomaly detection with coarse-knowledge-aware adversarial learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 16532–16540 (2025)
- [22] Iqbal, E., Khan, S.U., Javed, S., Moyo, B., Zweiri, Y., Abdulrahman, Y.: Multi-scale feature reconstruction network for industrial anomaly detection. *Knowledge-Based Systems* **305**, 112650 (2024)
- [23] Zavrtanik, V., Kristan, M., Skočaj, D.: Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8330–8339 (2021)
- [24] Li, C.-L., Sohn, K., Yoon, J., Pfister, T.: 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 (2021)
- [25] Zhang, X., Xu, M., Zhou, X.: Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16699–16708 (2024)
- [26] Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: 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 (2019)
- [27] Mishra, P., Verk, R., Fornasier, D., Piciarelli, C., Foresti, G.L.: Vt-adl: A vision transformer network for image anomaly detection and localization. *arXiv preprint arXiv:2104.10036* (2021)
- [28] Zhang, W., Lu, K., Wang, J., Liang, H., Gao, C., Xue, J.: Improving unsupervised ultrasonic image anomaly detection via frequency-spatial feature filtering and gaussian mixture modeling. *IEEE Transactions on Image Processing* **35**, 1567–1581 (2026) <https://doi.org/10.1109/TIP.2026.3659292>

- [29] Lyu, S., Mo, D., Wong, W.: Reb: Reducing biases in representation for industrial anomaly detection. *Knowledge-Based Systems* **290**, 111563 (2024)
- [30] Im, J., Son, Y., Hong, J.H.: Fun-ad: Fully unsupervised learning for anomaly detection with noisy training data. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 9447–9456 (2025). IEEE
- [31] Yang, H.-Y., Chen, H., Liu, L., Lin, Z., Chen, K., Wang, L., Han, J., Ding, G.: Context enhancement with reconstruction as sequence for unified unsupervised anomaly detection. *arXiv preprint arXiv:2409.06285* (2024)
- [32] Yao, X., Li, R., Qian, Z., Luo, Y., Zhang, C.: Focus the discrepancy: Intra-and inter-correlation learning for image anomaly detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6803–6813 (2023)
- [33] Yao, X., Li, R., Qian, Z., Wang, L., Zhang, C.: Hierarchical gaussian mixture normalizing flow modeling for unified anomaly detection. In: *European Conference on Computer Vision*, pp. 92–108 (2024). Springer
- [34] Rolih, B., Fučka, M., Skočaj, D.: Supersimplenet: Unifying unsupervised and supervised learning for fast and reliable surface defect detection. In: *International Conference on Pattern Recognition*, pp. 47–65 (2024). Springer
- [35] Liu, X., Wang, J., Leng, B., Zhang, S.: Unlocking the potential of reverse distillation for anomaly detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5640–5648 (2025)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [snarticle1.tex](#)