Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

HybridCaT: A Hybrid Convolution and Transformer Model for Multi-Class Unsupervised Anomaly Detection

Yunxin Liu
Shandong Women's University
Yifan Wang
Shandong Women's University
Danqing Liu
Chengdu University of Technology
Qian Yu
Shandong Women's University
Guanliang Wan
Qinghai Normal University
Yanhui Guo
guoyanhui@sdwu.edu.cn
Shandong Women's University

Research Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11780204/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

HybridCaT: A Hybrid Convolution and Transformer Model for Multi-Class Unsupervised Anomaly Detection

Yunxin Liu[1], Yifan Wang[1], Danqing Liu[2], Qian Yu[1], Guanliang Wan[3], Yanhui Guo[1*]

[1*]School of Artificial Intelligence, Shandong Women's University, Jinan, 250300, Shandong, China.
[2]College of Computer Science and Cyber Security (Pilot Software College), Chengdu University of Technology, Chengdu, 610059, Sichuan, China.
[3]School of Computer Science, Qinghai Normal University, Xining, 810000, Qinghai, China.

*Corresponding author(s). E-mail(s): guoyanhui@sdwu.edu.cn;
Contributing authors: 202333331047@stu.qhnu.edu.cn; w18379178@gmail.com; liudanqing@cdut.edu.cn; yuqian@sdwu.edu.cn; 1294621691@qq.com;

Abstract

Multi-class unsupervised anomaly detection plays a critical role in industrial visual inspection, where the core objective is to identify defects as deviations from normal patterns under the assumption that models are trained exclusively on normal samples. Owing to substantial variations among industrial products in terms of structural configurations, texture characteristics, and anomaly scales, effective detection requires a model capable of simultaneously capturing fine-grained local features and long-range global dependencies. To address these challenges, this paper proposes a hybrid feature reconstruction framework for multi-class unsupervised anomaly detection, termed HybridCaT. By leveraging the complementary strengths of convolutional neural networks and Transformer architectures, HybridCaT introduces a Parallel Dilated-Transformer (PD-Transformer) decoder to jointly model local texture details and global semantic representations, thereby enhancing the reconstruction of multi-scale normal feature distributions. Furthermore, a Dual-level Importance Feature Weighting (DIFW) module is designed to achieve dynamic cross-level feature fusion through intra-feature and inter-feature attention mechanisms. Extensive experiments on four public anomaly detection benchmarks—MVTec-AD, VisA, Real-IAD, and MVTec-AD2—demonstrate the strong and competitive performance of HybridCaT in both image-level anomaly detection and pixel-level localization.

Keywords: Multi-class anomaly detection, Convolutional neural networks, Vision Transformer, Feature reconstruction

1 Introduction

Anomaly Detection (AD) aims to identify data instances that significantly deviate from the dominant distribution of normal samples. With the rapid advancement of deep learning, anomaly detection techniques have been widely deployed across various application domains [1–3]. In industrial manufacturing [4], automated AD systems

can effectively replace manual inspection by enabling rapid and reliable identification of defective products, thereby reducing labor costs and minimizing human-induced errors.

However, industrial anomalies often exhibit substantial diversity in spatial scale, texture characteristics, and semantic complexity. Fine-grained defects such as surface scratches or tiny bubbles typically occupy highly localized regions, whereas structural defects such as contamination, fractures, or misalignments extend over much larger spatial extents. This pronounced multi-scale heterogeneity imposes a fundamental challenge: a unified detection model must simultaneously capture fine-grained local details and long-range global dependencies to achieve high-precision industrial anomaly detection.

Convolutional neural networks (CNNs) are effective in capturing local textures and fine-grained defect patterns, but their limited receptive fields hinder the modeling of long-range dependencies and global structural relationships. In contrast, Transformer-based models capture global contextual information through self-attention mechanisms, yet may be less sensitive to subtle local anomalies. This imbalance between local detail extraction and global dependency modeling limits the ability of single-paradigm architectures to handle industrial anomalies with diverse spatial scales and structural complexity.

To overcome these limitations, we propose HybridCaT (Hybrid Convolution and Transformer), a hybrid feature reconstruction framework that integrates the complementary strengths of CNNs and Transformers. Specifically, a pretrained Vision Transformer (ViT) encoder extracts hierarchical feature representations, which are adaptively aggregated by a Dual-level Importance Feature Weighting (DIFW) module. Furthermore, a Parallel Dilated-Transformer (PD-Transformer) decoder combines multi-scale dilated convolutions and Transformer-based global modeling to reconstruct normal feature representations while preserving both local details and global structural information.

Specifically, HybridCaT adopts a pretrained Vision Transformer (ViT) encoder to extract multi-stage hierarchical representations, preserving shallow-layer texture details while capturing high-level semantic structures. To mitigate the semantic and receptive field discrepancies among these multi-stage features, we design a Dual-level Importance Feature Weighting (DIFW) module. DIFW performs adaptive feature aggregation via a dual-attention mechanism: the intra-feature attention enhances salient channel responses within individual feature maps, whereas the inter-feature attention dynamically models the relative importance across hierarchical feature levels. This coordinated weighting strategy facilitates effective cross-scale information integration and yields more discriminative anomaly representations.

The main contributions of this work are summarized as follows:

1. We propose HybridCaT, a unified hybrid feature reconstruction framework that integrates CNNs and Transformers to capture multi-scale normal feature distributions, thereby enhancing sensitivity to heterogeneous industrial anomalies.
2. We design a DIFW module with a dual-attention mechanism to enable adaptive and cross-level feature fusion across hierarchical representations.
3. We develop a PD-Transformer decoder that captures multi-scale dependencies through parallel multi-scale dilated convolutions and Transformer branches, and further improves robustness via an integrated denoising mechanism.

2 Related work

Multi-class unsupervised anomaly detection (MUAD) aims to identify anomalous samples across multiple product categories using a unified model trained exclusively on normal data [5]. Compared with single-class anomaly detection methods [6, 7], MUAD must handle greater variations in object appearance, defect characteristics, and feature distributions, making the joint modeling of local details and global semantic information particularly important.

2.1 CNN- and Transformer-Based MUAD

CNN-based methods leverage local convolutional operations to capture fine-grained textures and structural patterns. Representative approaches include OmniAL [8], SNL [9], OMAC [10], and

InvAD [11], which explore anomaly synthesis, feature distillation, memory-based detection, and feature reconstruction, respectively. Recent studies further investigate domain-invariant representation learning and long-tailed anomaly detection through VLM-guided augmentation, vector quantization, and semantic feature modeling [12, 13].

Transformer-based methods employ self-attention to capture long-range dependencies and global contextual relationships. Representative approaches include PNPT [14], ViTAD [15], MSTAD [16], and HVQ-Trans [17]. More recent studies address issues such as anomaly leakage, inter-class interference, and computational efficiency through collaborative reconstruction, feature decomposition, class-aware representations, and mixture-of-experts architectures [18–21].

Despite their effectiveness, CNN-based models are limited in capturing long-range dependencies, whereas Transformer-based approaches may lack sufficient sensitivity to fine-grained local patterns. These limitations motivate hybrid architectures that integrate complementary local and global representations for multi-scale anomaly detection.

2.2 Hybrid CNN-Transformer Architectures

Hybrid architectures generally adopt sequential or parallel designs to combine convolutional and Transformer representations. Sequential approaches, such as DETR [22], LeViT [23], and CoAtNet [24], integrate convolutional operations and Transformer blocks in a cascaded manner. In contrast, parallel architectures, including Conformer [25], Mobile-Former [26], and CDDFuse [27], employ separate branches to capture complementary local and global information.

Different from these existing architectures, HybridCaT is specifically designed for multi-class industrial anomaly detection. It introduces a parallel decoding structure that integrates multi-scale dilated convolutions with Transformer-based global modeling, enabling the simultaneous reconstruction of fine-grained local details and long-range contextual representations.

3 Methodology

3.1 Model Overall Structure

As illustrated in Fig.1, HybridCaT follows an encoder-decoder architecture. A pretrained DINOv2 Vision Transformer (ViT) [28] serves as the encoder to extract hierarchical feature representations, which are adaptively aggregated by the proposed DIFW module. The decoder consists of three cascaded PD-Transformer stages, each integrating parallel multi-scale dilated convolutional branches and a Transformer module. At each stage, five intermediate representations are dynamically fused and progressively forwarded to the subsequent stage, enabling multi-scale feature reconstruction.

3.2 Feature Encoder

Features extracted from different encoder stages encode varying semantic granularity and receptive field coverage. Shallow features primarily preserve low-level spatial details such as edges and textures, whereas deeper features progressively encode high-level semantic structures and global contextual information. Leveraging this hierarchical property is essential for modeling anomalies with diverse spatial scales and semantic complexity.

Accordingly, we extract features from four distinct stages of the DINOv2 encoder to fully exploit its hierarchical representational capacity. The 12-layer ViT encoder is partitioned into four stages, each consisting of three Transformer layers. The extracted multi-stage features are denoted as $F = \{F_1, F_2, F_3, F_4\}$, where $F_i \in \mathbb{R}^{L \times C}$.

3.3 Feature Fusion

Multi-stage features contain complementary information but differ in semantic levels and receptive fields. Direct aggregation may introduce redundant responses due to their unequal importance.

To address this issue, we propose a Dual-Level Importance Feature Weighting (DIFW) module, which jointly models intra-feature channel saliency and inter-feature importance across encoder stages to achieve adaptive feature aggregation.

As illustrated in Fig.2, DIFW takes feature maps from four encoder stages as input and

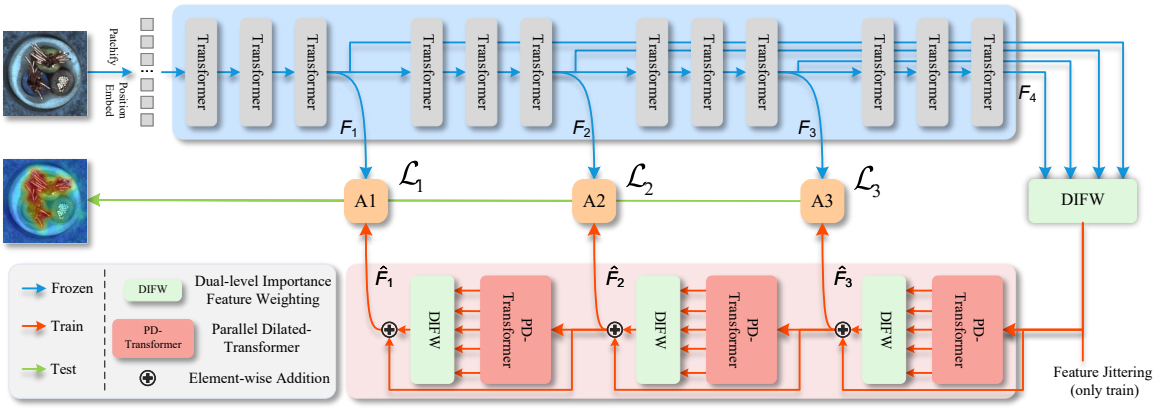


Fig. 1 HybridCaT Framework: HybridCaT first processes the input image using patch partitioning and positional embedding, then feeds it into a pre-trained ViT encoder to extract multi-level feature representations. Next, these features are integrated via the DIFW module, and Gaussian noise is added to enhance robustness before passing the fused features into the decoder. The decoder, utilizing the PD-Transformer architecture, models multi-scale dependencies, followed by a second feature fusion stage through the DIFW module, ultimately generating the reconstructed features. Multi-scale comparisons between F_i and $\hat{F}_i$ yield the precise anomaly heatmap A_i

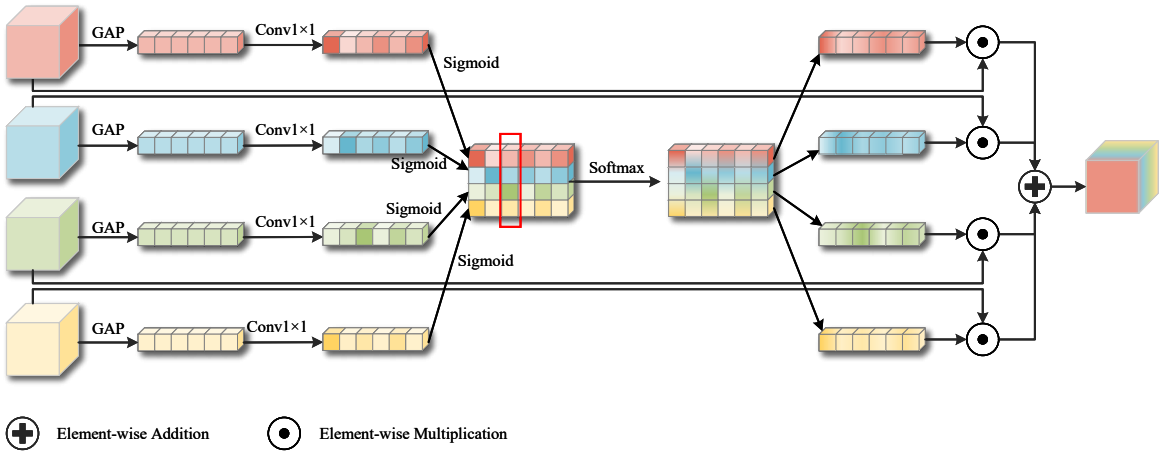


Fig. 2 Dual-level Importance Feature Weighting. It adaptively fuses multi-stage hierarchical features by jointly modeling intra-feature channel saliency and inter-stage competition to generate dynamic importance weights

produces a unified fused representation for the decoder. For each feature map F_i, a compact channel descriptor is extracted via global average pooling followed by pointwise projection:

$$y_i = \mathrm{Conv}_{1\times1}\left(\mathrm{GAP}(F_i)\right). \tag{1}$$

The descriptor provides global channel statistics for estimating feature importance.

Sigmoid activation models intra-feature channel importance, while Softmax normalization introduces competition among encoder stages for each channel. The resulting dual-level weighting is formulated as:

$$W = \mathrm{Softmax}(\mathrm{Concat}(\sigma(y_i), \mathrm{dim} = 2)). \tag{2}$$

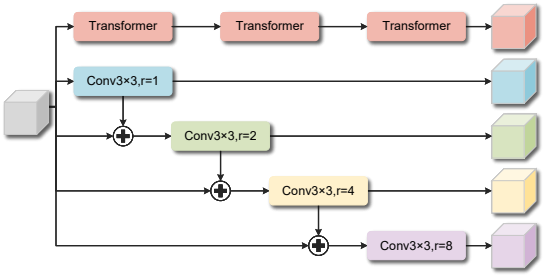


Fig. 3 Parallel Dilated-Transformer. To overcome the representational biases of a single paradigm, it explicitly balances global structural modeling (via stacked Transformer layers) and fine-grained local extraction (via parallel dilated convolutions)

Finally, the learned importance weights are applied to the corresponding feature maps and aggregated to obtain the fused representation:

$$F_f = \sum F_i \cdot W_i. \tag{3}$$

3.4 Feature Decoder

The decoder aims to reconstruct discriminative representations from the fused feature F_f while preserving both global structural coherence and local spatial details. However, relying on a single modeling paradigm is often insufficient: Transformer-based decoders excel at capturing long-range dependencies but may overlook fine-grained spatial patterns, whereas convolutional operations are effective for local context modeling yet lack global awareness. Motivated by this complementarity, we design a Parallel Dilated-Transformer decoder (PD-Transformer) to jointly integrate global and local feature modeling within a unified reconstruction framework.

As illustrated in Fig.3, the PD-Transformer comprises two parallel branches: a Transformer branch for global dependency modeling and a dilated convolution branch for multi-scale local context extraction. To enhance robustness and encourage learning of normal feature distributions, Gaussian noise is injected into the fused feature F_f, following the denoising strategy in UniAD [5]. A learnable positional embedding is further added to preserve spatial correspondence for Transformer processing.

Given the noise-perturbed and position-embedded feature $F_f \in \mathbb{R}^{L\times C}$, the Transformer branch applies three stacked Transformer layers with standard self-attention to model long-range dependencies, yielding a global representation $F_f^1 \in \mathbb{R}^{L\times C}$.

In parallel, the convolution branch reshapes $F_f \in \mathbb{R}^{L\times C}$ into spatial form $[C, H, W]$ with $H = \sqrt{L}$. A series of dilated convolutions with increasing dilation rates are applied to progressively enlarge the receptive field:

$$F_f^2 = Conv2D_{k=3,r=1}(F_f), \tag{4}$$
$$F_f^3 = Conv2D_{k=3,r=2}(F_f + F_f^2), \tag{5}$$
$$F_f^4 = Conv2D_{k=3,r=4}(F_f + F_f^3), \tag{6}$$
$$F_f^5 = Conv2D_{k=3,r=8}(F_f + F_f^4), \tag{7}$$

Each convolution block consists of depthwise convolution, instance normalization, and SiLU activation, enabling efficient multi-scale context aggregation.

The global feature F_f^1 is reshaped into spatial form and, together with the multi-scale local features $F_f^2 - F_f^5$, are subsequently fused using the DIFW module to obtain the unified representation F_{fused}.

Finally, to preserve the original fused information and stabilize reconstruction, a residual reconstruction strategy is adopted. The decoder iteratively refines the reconstructed features by adding the fused representation:

$$\hat{F}_i = \hat{F}_{i-1} + F_{fused}, \quad i = 1, 2, 3, \tag{8}$$

where $\hat{F}_0 = F_f$. This iterative residual design facilitates progressive refinement while preventing excessive deviation from the original fused feature.

By jointly integrating global Transformer-based dependency modeling and multi-scale dilated convolutional feature extraction in a parallel manner, the proposed PD-Transformer decoder achieves balanced reconstruction with strong global structural consistency and fine-grained local discriminability.

3.5 Loss Function and Anomaly Map

During training, reconstruction quality is optimized by maximizing cosine similarity between

encoder features and their reconstructed counterparts. The cosine similarity loss is defined as:

$$\mathcal{L}_{cos}(F, \hat{F}) = 1 - \frac{F \cdot \hat{F}}{\| F \|_2 \| \hat{F} \|_2}. \tag{9}$$

Here, F and $\hat{F}$ denote the encoder and decoder features, respectively. The loss is computed at each stage and summed to obtain the total objective:

$$\mathcal{L}_{total} = \sum_{i=1}^{n} \mathcal{L}_{cos}(F_i, \hat{F}_i). \tag{10}$$

Where n is the number of stages.

During inference, cosine similarity maps from all stages are similarly aggregated to generate the final anomaly map, where lower similarity scores indicate higher anomaly likelihood.

4 Experiments and results

4.1 Dataset Descriptions

Experiments are conducted on four public industrial anomaly detection benchmarks. MVTec-AD [1] contains 15 object and texture categories with 73 defect types and 5,354 images. VisA [29] comprises 12 object categories with diverse appearances, complex structures, and varying defect scales. Real-IAD [30] is a large-scale multi-view dataset containing 30 object categories, providing realistic industrial inspection scenarios. MVTec-AD2 [31] introduces challenging conditions, including transparent and overlapping objects, complex illumination, high intra-class variability, and extremely small defects. For MVTec-AD2, the official training and test splits are used for model optimization and evaluation, respectively.

4.2 Evaluation Metrics

We employ four widely adopted metrics to evaluate anomaly detection and localization performance: Area Under the Receiver Operating Characteristic Curve (AUROC), Average Precision (AP), F1 score under the optimal threshold (F1-max), and Intersection over Union under the optimal threshold (IoU-max).

4.3 Implementation Details

All input images are resized to 256 × 256. The model is optimized using AdamW with an initial learning rate of 1e-4 and a decay rate of 1e-4. Training is performed for 500 epochs with a batch size of 8. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU with 24 GB memory.

4.4 Comparison with SOTA MUAD Methods

4.4.1 Comparison Methods

Six representative MUAD methods are selected for comparison, including UniAD [5], ViTAD [15], HVQ-Trans [17], DiAD [32], GLASS [33], and CRAS [34]. These methods cover reconstruction-based and embedding-based anomaly detection paradigms, including Transformer reconstruction, hierarchical feature modeling, diffusion-based reconstruction, and anomaly synthesis. For fair comparison, GLASS is adapted to the multi-class setting, and all methods are evaluated under identical experimental settings.

4.4.2 Quantitative Evaluations

Comprehensive quantitative comparisons are conducted at both the image level and pixel level to evaluate anomaly detection accuracy and localization capability.

MVTec-AD. The experimental results are shown in Tables 1 and 2. HybridCaT demonstrates outstanding performance in both image-level and pixel-level tasks. For the image-level task, HybridCaT achieves average mAUROC, mAP, and mF1-max scores of 99.3%, 99.7%, and 98.3%, respectively, attaining optimal or competitive performance on 12 out of 15 categories, with particularly notable results in categories such as Metal Nut, Grid, and Cable. In the pixel-level task, HybridCaT achieves average mAUROC, mAP, mF1-max, and mIoU-max scores of 98.3%, 57.9%, 61.9%, and 45.7% respectively, showing significant advantages on categories such as Screw, Tile, and Bottle. These results indicate that HybridCaT effectively captures both global structural information and fine-grained local textures, enabling accurate localization of subtle defects and complex anomaly patterns.

Table 1 Image-level multi-class anomaly classification results with mAU-ROC/mAP/mF1-max metrics on MVTec-AD.

	Category/Method	Embedding-based			Reconstruction-based			
		GLASS[33] (ECCV'24)	CRAS[34] (TII'25)	UniAD[5] (NeurIPS'22)	HVQ-Trans[17] (NeurIPS'23)	DiAD[32] (AAAI'24)	ViTAD[15] (CVIU'25)	HybridCaT (Ours)
Texture	Carpet	99.0/99.7/98.3	98.6/99.6/97.2	99.8/**99.9**/**99.4**	**99.9**/**99.9**/**99.4**	99.4/**99.9**/98.3	99.5/**99.9**/**99.4**	99.6/**99.9**/98.9
	Grid	83.0/93.9/86.8	98.3/99.5/99.1	99.3/99.8/99.1	94.4/98.1/93.1	98.5/99.8/97.7	99.7/99.9/99.1	**100./100./100.**
	Leather	99.5/99.8/98.9	**100./100./100.**	**100./100./100.**	**100./100./100.**	99.8/99.7/97.6	**100./100./100.**	**100./100./100.**
	Tile	99.9/**100.**/99.4	99.1/99.7/98.8	99.9/99.9/99.4	99.3/99.7/98.2	96.8/99.9/98.4	**100./100./100.**	**100./100./100.**
	Wood	99.6/99.9/98.4	99.3/99.8/99.2	98.9/99.7/97.5	97.8/99.3/96.7	**99.7**/**100./100.**	98.7/99.6/96.7	98.9/99.7/97.4
Object	Bottle	98.6/99.5/97.6	**100./100./100.**	**100./100./100.**	**100./100./100.**	99.7/96.5/91.8	**100./100./100.**	**100./100./100.**
	Cable	86.9/92.0/84.7	98.2/98.9/94.6	94.8/97.0/90.7	97.8/98.7/94.1	94.8/98.8/95.2	98.5/99.1/95.7	**99.7**/**99.8**/**97.8**
	Capsule	82.6/94.9/94.3	95.9/99.0/96.4	93.7/98.4/96.3	93.7/98.4/**96.5**	89.0/97.5/95.5	95.4/99.0/95.5	**97.4**/**99.5**/96.3
	Hazelnut	99.7/99.8/98.6	**100./100./100.**	**100./100./100.**	99.9/99.9/99.3	99.5/99.7/97.3	99.8/99.9/98.6	**100./100./100.**
	Metal Nut	95.0/98.9/95.0	99.7/99.9/98.9	98.3/99.5/98.4	99.9/99.9/99.5	99.1/96.0/91.6	99.7/99.9/98.4	**100./100./100.**
	Pill	84.4/96.7/93.1	96.2/99.3/96.5	94.4/99.0/95.4	95.7/99.2/96.1	95.7/98.5/94.5	96.2/99.3/96.4	**98.3**/**99.7**/**97.9**
	Screw	78.4/91.7/87.1	85.0/95.1/86.5	95.3/98.5/92.9	94.6/98.0/93.8	90.7/**99.7**/**97.9**	91.3/97.0/93.0	**97.2**/99.1/95.1
	Toothbrush	89.4/96.1/91.8	**100./100./100.**	89.7/95.3/95.2	93.9/97.3/95.2	99.7/99.9/99.2	98.9/99.6/96.8	**100./100./100.**
	Transistor	86.8/84.4/81.0	98.9/96.9/**97.6**	99.8/**99.8**/97.1	**99.9**/**99.8**/**97.6**	99.8/99.6/97.4	98.8/98.3/92.5	98.5/98.0/91.9
	Zipper	97.8/99.4/97.5	**99.8**/**100./99.2**	98.6/99.6/97.1	97.0/99.1/97.5	95.1/99.1/94.4	97.6/99.3/97.3	99.7/99.9/**99.2**
	Average	92.1/96.4/93.5	97.9/99.2/97.6	97.5/99.1/97.3	97.6/99.1/97.1	97.2/99.0/96.5	98.3/99.4/97.3	**99.3**/**99.7**/**98.3**

The best results are highlighted in **bold**.

Table 2 Pixel-level multi-class anomaly classification results with mAU-ROC/mAP/mF1-max/mIoU-max metrics on MVTec-AD.

	Category/Method	Embedding-based			Reconstruction-based			
		GLASS[33] (ECCV'24)	CRAS[34] (TII'25)	UniAD[5] (NeurIPS'22)	HVQ-Trans[17] (NeurIPS'23)	DiAD[32] (AAAI'24)	ViTAD[15] (CVIU'25)	HybridCaT (Ours)
Texture	Carpet	**99.2**/**72.6**/**67.0**/**50.3**	98.4/52.2/54.7/37.7	98.5/49.9/51.1/34.5	98.6/56.2/56.5/39.4	98.6/42.2/46.4/46.0	99.0/60.5/64.1/47.2	**99.2**/62.0/65.4/48.6
	Grid	89.7/17.9/27.9/16.2	98.1/21.2/28.2/16.4	63.1/ 0.7/ 1.9/17.7	96.8/24.2/30.0/17.5	96.6/**66.0**/**64.1**/19.1	98.6/31.2/36.7/22.7	**99.0**/35.4/40.4/**25.3**
	Leather	98.8/50.0/49.3/32.7	99.1/41.5/46.4/30.2	98.8/32.9/34.4/21.5	98.8/33.7/36.2/21.7	98.8/**56.1**/**62.3**/29.4	**99.6**/52.1/55.8/**40.3**	99.5/53.1/55.0/37.9
	Tile	**98.5**/**81.7**/**76.0**/**61.3**	96.3/70.6/64.7/47.8	91.8/42.1/50.6/33.2	92.5/43.4/53.8/35.0	92.4/65.7/64.1/41.6	96.6/56.4/68.8/52.5	97.1/58.2/70.8/54.8
	Wood	**96.8**/**64.0**/**59.3**/**42.1**	93.6/46.4/47.0/30.7	93.2/37.2/41.5/27.0	92.6/38.6/43.8/26.8	93.3/43.3/43.5/30.1	96.4/60.6/58.3/41.3	**96.8**/59.3/58.9/41.8
Object	Bottle	97.1/63.1/62.9/45.9	98.6/**82.1**/**77.5**/**63.3**	98.1/66.0/69.2/50.9	98.3/72.0/70.0/53.6	98.4/52.2/54.8/57.7	98.8/79.9/75.6/61.6	**99.0**/**80.6**/**77.5**/**63.3**
	Cable	93.0/36.5/42.8/27.2	98.2/**69.9**/**65.2**/**48.3**	97.3/39.9/45.2/30.4	97.9/51.1/56.5/42.9	96.8/50.1/57.8/35.9	96.2/43.1/47.4/30.2	**98.3**/55.2/62.5/45.5
	Capsule	97.4/29.8/37.6/23.1	**98.8**/**45.5**/**52.1**/**35.2**	98.5/42.7/46.5/30.0	**98.8**/44.2/48.9/33.4	97.1/42.0/45.3/30.0	98.3/42.7/47.8/31.9	98.4/44.5/48.0/31.6
	Hazelnut	98.0/50.2/54.6/37.5	98.3/57.1/58.9/41.8	98.1/55.2/56.8/39.2	98.8/63.3/62.3/45.0	98.3/**79.2**/**80.4**/44.2	99.0/64.6/64.0/47.9	**99.1**/67.1/67.3/50.8
	Metal Nut	91.5/58.6/58.8/41.7	**98.5**/**92.5**/**84.0**/**72.3**	62.7/14.6/29.2/47.7	96.6/68.2/73.2/58.2	97.3/30.0/38.3/67.4	96.4/75.1/77.3/62.9	97.4/75.5/80.7/67.7
	Pill	96.8/73.2/68.6/52.3	98.6/**84.9**/**75.6**/**60.8**	95.0/44.0/53.9/37.0	96.4/51.4/56.4/38.9	95.7/46.0/51.4/39.3	**98.7**/77.8/75.2/60.6	98.6/73.2/74.2/58.9
	Screw	98.0/27.6/32.4/19.3	97.5/28.5/35.3/21.5	98.3/28.7/37.6/26.9	98.9/28.4/35.6/20.2	97.9/**60.6**/**59.6**/23.9	99.0/34.0/41.0/24.2	**99.4**/43.5/45.4/**29.4**
	Toothbrush	98.6/61.3/59.7/42.5	99.0/61.8/64.6/47.8	98.4/34.9/45.7/32.3	98.5/38.9/49.7/34.2	99.0/**78.7**/**72.8**/**49.2**	**99.1**/51.3/61.9/43.9	99.0/44.1/60.5/43.4
	Transistor	86.8/37.9/39.0/24.2	97.1/70.1/65.2/48.4	97.9/59.5/64.6/46.7	**98.5**/**74.7**/**71.7**/**61.6**	95.1/15.6/31.7/46.2	93.9/58.4/55.3/37.9	95.5/63.2/60.3/43.1
	Zipper	**98.9**/67.9/65.0/48.2	98.7/**70.1**/**65.3**/**48.5**	96.8/40.1/49.9/33.5	97.5/39.3/48.6/30.9	96.2/30.7/30.0/34.3	95.9/42.6/50.8/32.6	97.7/53.4/60.8/43.7
	Average	95.9/52.8/53.4/37.6	97.9/**59.6**/59.0/43.4	96.8/43.4/49.5/33.9	97.3/48.5/52.9/37.3	96.8/52.6/55.5/39.6	97.7/55.3/58.7/42.5	**98.3**/57.9/**61.9**/**45.7**

VisA. The results on VisA are summarized in Tables 3. HybridCaT achieves 94.2% mAUROC, 95.2% mAP, and 90.6% mF1-max at the image level. Pixel-level performance remains strong with 98.7% mAUROC and competitive mAP, mF1-max, and mIoU-max. The experimental results indicate that HybridCaT is capable of effectively addressing anomaly detection and localization tasks in complex industrial scenarios.

Real-IAD. The quantitative results on Real-IAD are reported in Table 3. At the image level, HybridCaT achieves average mAUROC, mAP, and mF1-max scores of 86.7%, 85.1%, and 77.5%, respectively, outperforming all competing methods. This demonstrates strong robustness and generalization capability under real-world industrial conditions characterized by diverse materials, viewpoints, and defect patterns. For the pixel-level task, HybridCaT attains an mAUROC of 97.3%, with scores of mAP, mF1-max, and mIoU-max of 29.9%, 37.1% and 23.2%, respectively, showcasing its potential in local anomaly localization. The experimental results indicate that HybridCaT has broad application prospects in real industrial environments.

MVTec-AD2. The results on the challenging MVTec-AD2 benchmark are also summarized in Table 3. HybridCaT achieves the highest image-level mAUROC of 68.4%, along with competitive mAP and mF1-max scores. At the pixel level, HybridCaT consistently achieves the best performance across all metrics. Notably, MVTec-AD2 includes transparent and overlapping objects,

Table 3 Image/Pixel-level multi-class anomaly classification results on VisA, Real-IAD and MVTec-AD2.

Dataset		Metrics	Embedding-based		Reconstruction-based				
			GLASS[33] (ECCV'24)	CRAS[34] (TII'25)	UniAD[5] (NeurIPS'22)	HVQ-Trans[17] (NeurIPS'23)	DiAD[32] (AAAI'24)	ViTAD[15] (CVIU'25)	HybridCaT (Ours)
VisA	Image	mAU-ROC	91.1	93.2	92.1	90.0	86.8	90.5	**94.2**
		mAP	93.6	94.9	93.6	91.7	88.3	91.7	**95.2**
		mF1-max	86.8	90.4	88.1	87.0	85.1	86.3	**90.6**
	Pixel	mAU-ROC	97.0	97.7	98.5	98.4	96.0	98.2	**98.7**
		mAP	**44.6**	39.1	35.0	32.2	26.1	36.6	41.5
		mF1-max	**47.9**	42.1	40.3	36.7	33.0	41.1	44.9
		mIoU-max	**32.2**	29.2	26.4	23.0	22.7	27.6	30.8
Real-IAD	Image	mAU-ROC	-	78.9	84.3	84.1	75.6	82.7	**86.7**
		mAP	-	77.7	82.4	81.1	66.4	80.1	**85.1**
		mF1-max	-	70.2	75.4	74.8	69.9	73.7	**77.5**
	Pixel	mAU-ROC	-	96.5	97.3	**97.5**	88.0	97.3	97.3
		mAP	-	20.7	24.5	21.2	2.9	24.2	**29.9**
		mF1-max	-	28.4	32.2	29.1	7.1	32.2	**37.1**
		mIoU-max	-	16.9	19.5	17.4	-	20.2	**23.2**
MVTec-AD2	Image	mAU-ROC	-	65.6	59.6	67.2	-	67.8	**68.4**
		mAP	-	**77.5**	68.9	75.4	-	74.3	74.8
		mF1-max	-	80.6	75.9	**80.9**	-	78.6	79.2
	Pixel	mAU-ROC	-	83.7	82.2	81.9	-	82.1	**83.9**
		mAP	-	18.3	10.2	13.6	-	19.2	**20.5**
		mF1-max	-	20.8	15.4	16.8	-	23.6	**24.6**
		mIoU-max	-	13.6	9.1	9.9	-	15.1	**15.6**

complex illumination setups, high intra-class variability, and extremely small defects, which significantly increase detection difficulty. The superior performance of HybridCaT under these conditions indicates that the proposed hybrid modeling strategy effectively captures both subtle local cues and global structural context.

4.4.3 Qualitative Evaluations

Fig.4 presents qualitative comparisons on MVTec-AD, VisA, and MVTec-AD2. Compared with existing methods, which may produce incomplete localization or false activations, HybridCaT generates more spatially coherent anomaly maps with clearer boundaries across different defect scales. It effectively localizes both small-scale defects, such as those in screws and capsules, and large-scale structural anomalies, demonstrating its ability to capture diverse industrial anomaly patterns.

4.5 Ablation Study

Three ablation studies are conducted on MVTec-AD to investigate the contributions of multi-stage encoder features, core modules, and Gaussian noise.

Table 4 The Impact of Multi-Stage Features of Encoder on Reconstruction Effectiveness.

	Image-level	Pixel-level
F_1	98.1	**98.3**
F_2	99.1	**98.3**
F_3	99.0	98.2
F_4	98.9	98.2
F_1, F_2, F_3, F_4	**99.3**	**98.3**

4.5.1 The Impact of Multi-Stage Features of Encoder on Reconstruction Performance

We divide the 12-layer Transformer-based ViT encoder into four stages (each with 3 layers) and extract features from each stage to evaluate their individual contributions to reconstruction performance. Five setups are compared: using features from each individual stage (F_1 to F_4) and fusing features from all four stages.

As shown in Table 4, the second-stage features yield the highest image-level performance with a score of 99.1%, while the first-stage features are more effective for pixel-level localization with 98.3%. Features from the deeper third and fourth stages also contribute positively, although

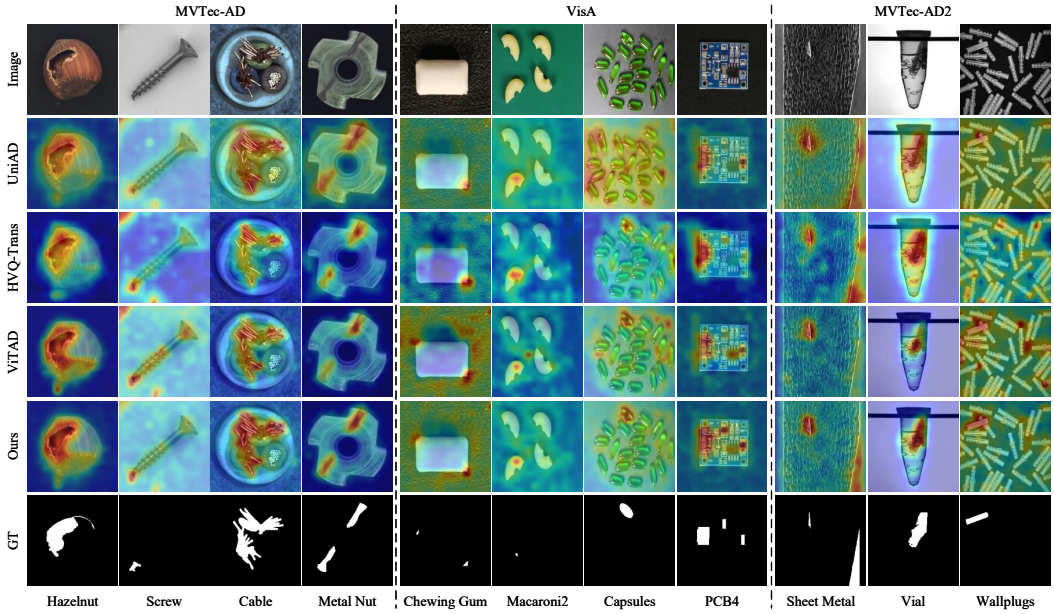


Fig. 4 Qualitative comparison of anomaly localization results on MVTec-AD, VisA, and MVTec-AD2. HybridCaT produces more coherent anomaly maps with clearer boundaries across different defect scales

their impact is less pronounced. When all stages are fused, the model attains the best image-level performance of 99.3%, whereas the improvement on pixel-level metrics remains marginal. These findings indicate that features at different depths serve complementary roles: mid-level features are more advantageous for global anomaly detection, whereas shallow features are better suited for precise local localization. This observation motivates the necessity of adaptive multi-scale feature fusion to simultaneously optimize detection accuracy and localization precision.

Table 5 Core Module Ablation Study.

	Image-level	Pixel-level
Only Transformer	98.3	97.7
PD-Transformer + Add	99.0	98.2
PD-Transformer + Concat	99.0	98.1
PD-Transformer + DIFW	**99.3**	**98.3**

4.5.2 Core Module Ablation Study

To evaluate the contribution of the proposed core modules, we first compare a baseline Transformer decoder with a decoder enhanced by the proposed PD-Transformer. Subsequently, different feature fusion strategiesare evaluated, including Add, Concat, and the proposed DIFW. Quantitative results are summarized in Table 5. Introducing the PD-Transformer yields a substantial improvement in both image-level and pixel-level performance, validating its effectiveness in jointly modeling global dependencies and local structures. Replacing conventional fusion strategies with DIFW further improves performance and consistently outperforms Add and Concat. The full model achieves the best overall results, reaching 99.3% image-level mAUROC and 98.3% pixel-level mAUROC. These findings demonstrate that the PD-Transformer strengthens local feature extraction, while the DIFW enables more effective multi-scale feature fusion. Their synergy significantly strengthens the reconstruction capability and anomaly detection performance of the proposed framework.

4.5.3 Ablation Study on Gaussian Noise Scale

To verify the effectiveness of introducing Gaussian noise and investigate the stability of the

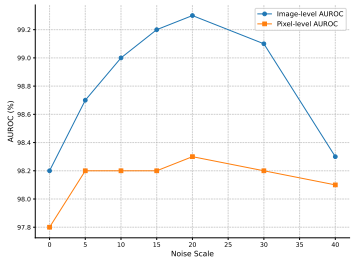


Fig. 5 Impact of different Gaussian noise scales on model robustness. The results indicate that a moderate noise scale of 20 provides the optimal trade-off between feature perturbation and reconstruction fidelity, achieving the highest image-level and pixel-level AUROC

model under different perturbation levels, we conducted ablation experiments with noise scales $\{0, 5, 10, 15, 20, 30, 40\}$. The results are presented in Fig. 5. Compared with the baseline, the introduction of Gaussian noise consistently enhances the performance, which demonstrates its positive contribution to model robustness. As the noise scale increases from 5 to 20, both image-level and pixel-level performance exhibit steady improvements, reaching the peak at a noise scale of 20, where the image-level AUROC achieves 99.3% and the pixel-level AUROC reaches 98.3%. When the noise scale further increases to 30 and 40, performance slightly degrades, indicating that excessive perturbations may distort the underlying feature distribution and adversely affect reconstruction accuracy. Overall, these results confirm that moderate noise injection effectively enhances generalization and robustness, and that a noise scale around 20 provides an optimal trade-off between robustness and reconstruction fidelity.

5 Conclusion

This paper presents HybridCaT, a hybrid feature reconstruction framework for multi-class unsupervised anomaly detection. By integrating a Dual-level Importance Feature Weighting (DIFW) module and a Parallel Dilated-Transformer (PD-Transformer) decoder, the proposed method adaptively fuses hierarchical features and jointly captures multi-scale local patterns and global contextual dependencies. Extensive experiments on four public benchmarks, including MVTec-AD, VisA, Real-IAD, and MVTec-AD2, demonstrate its competitive performance in image-level anomaly detection and pixel-level localization. The results validate the effectiveness of the proposed hybrid architecture and its potential for industrial anomaly detection.

Statements and Declarations

- **Funding:** This work was supported in part by the Shandong Provincial Natural Science Foundation under Grant ZR2023MF110, ZR2023MF037 and the Open Project Program of Novel Software Technology under Grant KFKT2025B33.

- **Competing interests:** The authors declare that they have no competing interests.
- **Ethics approval and consent to participate:** Not applicable.
- **Consent for publication:** All authors have approved the manuscript and consent to its publication.
- **Data availability:** The datasets used in this study are publicly available from their respective official sources: MVTec AD [1] (https://www.mvtec.com/research-teaching/datasets/mvtec-ad), VisA [29] (https://github.com/amazon-science/spot-diff), Real-IAD [30] (https://realiad4ad.github.io/Real-IAD/), and MVTec AD 2 [31] (https://www.mvtec.com/research-teaching/datasets/mvtec-ad-2). Access is subject to the respective dataset licenses and conditions. No new datasets were collected or generated in this study.
- **Materials availability:** Not applicable.
- **Code availability:** The source code of the proposed HybridCaT method will be made publicly available upon acceptance of this manuscript.
- **Author contribution:** Yunxin Liu: Writing-original draft, Methodology, Conceptualization.Yinfan Wang: Visualization, Investigation. Danqing Liu: Project administration, Resources, Supervision. Qian Yu: Formal analysis, Validation. Guanliang Wan: Software, Methodology. Yanhui Guo: Conceptualization, Methodology, Writing-review & editing, Project administration

References

[1] Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad — a comprehensive

real-world dataset for unsupervised anomaly detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9584–9592 (2019). https://doi.org/10.1109/CVPR.2019.00982

[2] Guo, J., Lu, S., Jia, L., Zhang, W., Li, H.: Encoder-decoder contrast for unsupervised anomaly detection in medical images. IEEE Trans. Med. Imaging **43**(3), 1102–1112 (2024) https://doi.org/10.1109/TMI.2023.3327720

[3] Bao, J., Sun, H., Deng, H., He, Y., Zhang, Z., Li, X.: Bmad: Benchmarks for medical anomaly detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4042–4053 (2024). https://doi.org/10.1109/CVPRW63382.2024.00408

[4] Liu, J., Xie, G., Wang, J., Li, S., Wang, C., Zheng, F., Jin, Y.: Deep industrial image anomaly detection: A survey. Mach. Intell. Res. **21**(1), 104–135 (2024) https://doi.org/10.1007/s11633-023-1459-z

[5] 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)

[6] Liu, Z., Zhou, Y., Xu, Y., Wang, Z.: Simplenet: A simple network for image anomaly detection and localization. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20402–20411 (2023). https://doi.org/10.1109/CVPR52729.2023.01954

[7] Zhang, X., Li, S., Li, X., Huang, P., Shan, J., Chen, T.: Destseg: Segmentation guided denoising student-teacher for anomaly detection. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3914–3923 (2023). https://doi.org/10.1109/CVPR52729.2023.00381

[8] Zhao, Y.: Omnial: A unified cnn framework for unsupervised anomaly localization. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3924–3933 (2023). https://doi.org/10.1109/CVPR52729.2023.00382

[9] Deng, H., Li, X.: Structural teacher-student normality learning for multi-class anomaly detection and localization. arXiv preprint arXiv:2402.17091 (2024) https://doi.org/10.48550/arXiv.2402.17091

[10] Tan, P., Wong, W.K.: Unsupervised anomaly detection and localization with one model for all category. Knowl.-Based Syst. **289**, 111533 (2024) https://doi.org/10.1016/j.knosys.2024.111533

[11] Zhang, J., Wang, C., Li, X., Tian, G., Xue, Z., Liu, Y., Pang, G., Tao, D.: Learning feature inversion for multi-class anomaly detection under general-purpose coco-ad benchmark. arXiv preprint arXiv:2404.10760 (2024) https://doi.org/10.48550/arXiv.2404.10760

[12] Xing, C., Xu, Y., Xu, Y., Cui, L.: Learning invariant discriminative patterns for unified anomaly detection. In: Proceedings of the 33rd ACM International Conference on Multimedia. MM '25, pp. 7874–7882. Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3746027.3755179 . https://doi.org/10.1145/3746027.3755179

[13] Ho, C.-H., Peng, K.-C., Vasconcelos, N.: Long-tailed anomaly detection with learnable class names. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12435–12446 (2024). https://doi.org/10.1109/CVPR52733.2024.01182

[14] Yao, H., Cao, Y., Luo, W., Zhang, W., Yu, W., Shen, W.: Prior normality prompt transformer for multiclass industrial image anomaly detection. IEEE Trans. Industrial Informatics **20**(10), 11866–11876 (2024) https://doi.org/10.1109/TII.2024.3413322

[15] Zhang, J., Chen, X., Wang, Y., Wang, C., Liu, Y., Li, X., Yang, M.-H., Tao, D.: Exploring plain vit features for multi-class unsupervised visual anomaly detection. Comput. Vis.

Image Underst. **253**, 104308 (2025) https://doi.org/10.1016/j.cviu.2025.104308

[16] Kang, B., Zhong, Y., Sun, Z., Deng, L., Wang, M., Zhang, J.: Mstad: A masked subspace-like transformer for multi-class anomaly detection. Knowl.-Based Syst. **283**, 111186 (2024) https://doi.org/10.1016/j.knosys.2023.111186

[17] Lu, R., Wu, Y., Tian, L., Wang, D., Chen, B., Liu, X., Hu, R.: Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. Adv. Neural Inf. Process. Syst. **36**, 8487–8500 (2023)

[18] Wang, Q., Wang, H., Gao, S., Guo, J., Xiong, L., Li, J., Bai, D., Zhang, W.: Collaborative Reconstruction and Repair for Multi-class Industrial Anomaly Detection (2025). https://doi.org/10.48550/arXiv.2512.11401 . https://arxiv.org/abs/2512.11401

[19] Wang, X., Wang, X., Bai, H., Lim, E.G., Xiao, J.: Decad: Decoupling anomalies in latent space for multi-class unsupervised anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 21568–21577 (2025)

[20] Jiang, X., Chen, Y., Nie, Q., Liu, J., Liu, Y., Wang, C., Zheng, F.: Toward Multi-class Anomaly Detection: Exploring Class-aware Unified Model against Inter-class Interference (2024). https://doi.org/10.48550/arXiv.2403.14213 . https://arxiv.org/abs/2403.14213

[21] Meng, S., Meng, W., Zhou, Q., Li, S., Hou, W., He, S.: Moead: A parameter-efficient model for multi-class anomaly detection. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024, pp. 345–361. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-73013-9_20

[22] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) Computer Vision – ECCV 2020, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13

[23] Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M.: Levit: a vision transformer in convnet's clothing for faster inference. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12239–12249 (2021). https://doi.org/10.1109/ICCV48922.2021.01204

[24] Dai, Z., Liu, H., Le, Q.V., Tan, M.: Coatnet: Marrying convolution and attention for all data sizes. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems, vol. 34, pp. 3965–3977. Curran Associates, Inc., ??? (2021). https://doi.org/10.48550/arXiv.2106.04803

[25] Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., Ye, Q.: Conformer: Local features coupling global representations for visual recognition. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 357–366 (2021). https://doi.org/10.1109/ICCV48922.2021.00042

[26] Chen, Y., Dai, X., Chen, D., Liu, M., Dong, X., Yuan, L., Liu, Z.: Mobile-former: Bridging mobilenet and transformer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5260–5269 (2022). https://doi.org/10.1109/CVPR52688.2022.00520

[27] Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R., Van Gool, L.: Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5906–5916 (2023). https://doi.org/10.1109/CVPR52729.2023.00572

[28] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual

features without supervision. arXiv preprint arXiv:2304.07193 (2023)

[29] Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: European Conference on Computer Vision, pp. 392–408 (2022). https://doi.org/10.1007/978-3-031-20056-4_23 . Springer

[30] 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: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22883–22892 (2024). https://doi.org/10.1109/CVPR52733.2024.02159

[31] Heckler-Kram, L., Neudeck, J.-H., Scheler, U., König, R., Steger, C.: The mvtec ad 2 dataset: Advanced scenarios for unsupervised anomaly detection. Int. J. Comput. Vision **134**(4) (2026) https://doi.org/10.1007/s11263-026-02743-0

[32] 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). https://doi.org/10.1609/aaai.v38i8.28690

[33] Chen, Q., Luo, H., Lv, C., Zhang, Z.: A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024, pp. 37–54. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-72855-6_3

[34] Chen, Q., Luo, H., Yao, H., Luo, W., Qu, Z., Lv, C., Zhang, Z.: Center-aware residual anomaly synthesis for multiclass industrial anomaly detection. IEEE Trans. Ind. Inform., 1–11 (2025) https://doi.org/10.1109/TII.2025.3575122