One *Dinomaly2* Detect Them All: A Unified Framework for Full-Spectrum Unsupervised Anomaly Detection

Jia Guo, Shuai Lu, Lei Fan, Zelin Li, Donglin Di, Yang Song, Weihang Zhang, Wenbing Zhu, Hong Yan, *Life Fellow, IEEE*, Fang Chen, Huiqi Li, *Senior Member, IEEE*, Hongen Liao, *Senior Member, IEEE*

Abstract—Unsupervised anomaly detection (UAD) has evolved from building specialized single-class models to unified multi-class models, yet existing multi-class models significantly underperform the most advanced one-for-one counterparts. Moreover, the field has fragmented into specialized methods tailored to specific scenarios (multi-class, 3D, few-shot, etc.), creating deployment barriers and highlighting the need for a unified solution. In this paper, we present Dinomaly2, the first unified framework for full-spectrum image UAD, which bridges the performance gap in multi-class models while seamlessly extending across diverse data modalities and task settings. Guided by the "less is more" philosophy, we demonstrate that the orchestration of simple elements—universal representations, basic dropouts, simplified attention mechanisms, contextual recentering, and loosened optimization objectives—achieves superior performance in a standard reconstruction-based framework. This methodological minimalism enables natural extension across diverse tasks without modification, establishing that simplicity is the foundation of true universality. Extensive experiments on 12 UAD benchmarks demonstrate Dinomaly2's full-spectrum superiority across multiple *modalities* (2D, multi-view, RGB-3D, RGB-IR), *task settings* (single-class, multi-class, inference-unified multi-class, few-shot) and *application domains* (industrial, biological, outdoor). For example, our unified multi-class model achieves unprecedented 99.9% and 99.3% image-level (I-) AUROC on MVTec-AD and VisA respectively. For multi-view and multi-modal inspection, Dinomaly2 demonstrates state-of-the-art performance with minimum adaptations. Moreover, using only 8 normal examples per class, our method surpasses previous full-shot models, achieving 98.7% and 97.4% I-AUROC on MVTec-AD and VisA. The combination of minimalistic design, computational scalability, and universal applicability positions Dinomaly2 as a unified solution for the full spectrum of real-world anomaly detection applications.

Index Terms—Anomaly Detection, Unsupervised Learning, Multi-Modal, Few-shot, Unified Model

1 Introduction

Unsupervised anomaly detection (UAD) aims to identify anomalous patterns in data without prior knowledge of anomaly types, serving as a fundamental task in computer vision with applications ranging from industrial quality control [1], [2], medical diagnosis [3], [4], to surveillance systems [5], [6]. The inherent scarcity and diversity of anomalies make this task particularly challenging, as models must learn to represent normal patterns from limited data while remaining sensitive to deviations during inference.

Early image UAD approaches follow a one-class-one-model paradigm [7], [8], where separate models are trained for each object category or scene (Fig. 1(a.1)). While this strategy has proven effective in controlled scenarios, it faces significant challenges in real-world applications. The storage overhead, computational complexity, and maintenance burden of managing hundreds of individual models become prohibitive in practical deployments. Recent advances have shifted toward *multi-class UAD* (MUAD, Fig. 1(a.2)), where a single unified model is designed to handle multiple object categories simultaneously [9]–[11]. Recent methods leverage a variety of techniques, including neighbor-masked attention [9], synthetic anomalies [10], vector quantization [12], diffusion model [13], [14], and state space model (Mamba) [15] to capture compact representations of normal patterns, which are critical for distinguishing anomalies. However, when intra-class normal patterns become excessively complex due to the presence of diverse categories, the underlying distribution becomes difficult to characterize, consequently degrading detection performance [9]. Yet despite this proliferation of complex designs, there is still a non-negligible performance gap between the state-of-the-art (SOTA) multi-class and single-class models (Fig. 1(d)), which restricts the practical adoption of unified approaches

Furthermore, existing UAD approaches suffer from a more fundamental problem: *methodological fragmentation*. The field has evolved into a collection of specialized frameworks, each tailored to specific task settings. For example, methods designed for plain 2D inspection [9], [10] cannot be naturally extended to multi-modal data without modifications [16], [17]. Methods designed for few-shot UAD often require entirely new pipelines that leverage vision-language

Jia Guo, Donglin Di, and Hongen Liao are with Tsinghua University, Beijing, China (liao@tsinghua.edu.cn).
Shuai Lu, Weihang Zhang, and Huiqi Li are with Beijing Institute of Technology, Beijing, China (huiqili@bit.edu.cn).
Fang Chen is with Shanghai Jiao Tong University, Shanghai, China (chenfang@sjtu.edu.cn). Hongen Liao is also with this institution.
Zelin Li and Hong Yan are with City University of Hong Kong, Hong Kong SAR. Lei Fan and Yang Song are with the University of New South Wales, Sydney, Australia. Donglin Di is also with DZ Matrix. Wenbing Zhu is with Fudan University and Rongcheer Co., Ltd.
Corresponding authors: Fang Chen; Huiqi Li; Hongen Liao

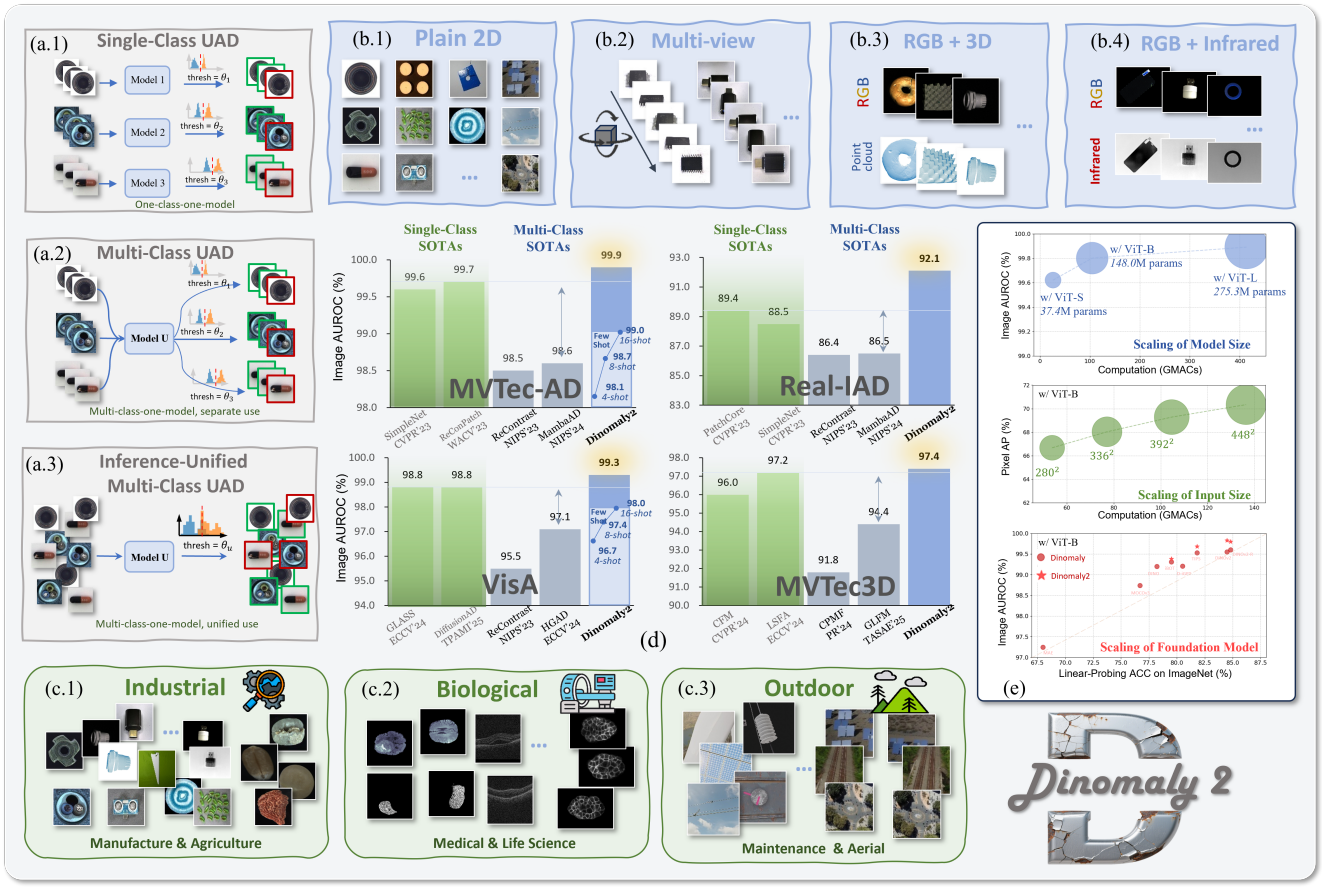

Fig. 1. Overview: task settings, benchmarks, and scalability of Dinomaly2. (a) Task settings include single-class UAD, multi-class UAD (MUAD), and inference-unified MUAD. (b) Multi-modal capabilities span 2D images, multi-view 2D, RGB+3D point cloud, and RGB+Infrared. (c) Cross-domain validation across industrial inspection, medical imaging, and aerial surveillance. (d) Dinomaly2 achieves performance on MVTec-AD (15 categories), VisA (12 categories), Real-IAD (30 categories×5 views), and MVTec3D (10 categories) that surpasses the corresponding state-of-the-art methods. (e) Scalability analysis shows consistent improvements with model size, input resolution, and foundation model quality (on MVTec-AD).

models or meta-learning [18], [19]. This proliferation of specialized methods creates huge barriers for practitioners. They must master multiple frameworks, maintain disparate codebases, and manage incompatible architectures when deploying anomaly detection across different scenarios. As UAD applications continue to diversify, the demand for unified frameworks become paramount. In this work, we aim to challenge the prevailing trend toward architectural complexity with a counterintuitive proposition: universal anomaly detection requires simplification, not specialization.

Here we present **Dinomaly2**, the first **multi-class, full-spectrum** UAD method that achieves unprecedented unification across data modalities, task settings, and application domains through principled minimalism. Built on the epistemic behavior [20] of neural networks, Dinomaly2 detects anomalous regions via reconstruction error (Fig. 2). The core of Dinomaly2 lies in the ***"less is more"*** philosophy, presenting five simple yet essential components:

- To begin with, we demonstrate that self-supervised Vision Transformers (ViTs), when properly scaled, provide universal feature representations that generalize across diverse UAD modalities and domains.
- We propose *Noisy Bottleneck* that activates the built-in Dropout in an MLP to prevent the network from over-generalization, as an alternative to meticulously designed synthetic anomaly and feature noise.
- We exploit *Unfocused Linear Attention*'s inherent inability to focus as a feature rather than a limitation, naturally preventing the propagation of identical information during reconstruction.
- We introduce *Context-Aware Recentering* through simple feature subtraction with class tokens, elegantly resolving the multi-class confusion where identical features may be normal in one context but anomalous in another.
- We propose *Loose Reconstruction* that deliberately relaxes layer-to-layer and point-by-point correspondence, preventing the decoder from perfectly mimicking the encoder.

Moreover, we present simple strategies to extend vanilla 2D Dinomaly2 to achieve SOTA results on diverse scenarios including multi-view inspection, RGB+3D point cloud, RGB+Infrared (IR), and few-shot UAD (Fig. 1(b.2-4)). We further advance the field by introducing the *inference-unified MUAD* setting (Fig. 1(a.3)), where anomalies across mixed categories must be detected with a single threshold—a crucial step toward truly unified deployment. This setting exposes the limitations of existing methods that rely on category-specific calibration while demonstrating Dinomaly2's robustness in this challenging setting.

We validate Dinomaly2's universality through the most comprehensive evaluation, spanning 12 benchmark datasets

(147 distinct categories/scenes, Table 1) across *four data modalities* (2D, multi-view 2D, RGB-3D, RGB-IR), *four task settings* (single-class, multi-class, inference-unified multi-class, few-shot), and *three application domains* (industrial, biological, outdoor). On popular 2D industrial datasets, our unified model achieves 99.9%, 99.3%, and 92.1% image-level (I-) AUROC on MVTec-AD [1], VisA [21], and Real-IAD [2] respectively, surpassing both multi-class and single-class SOTAs as shown in Fig. 1(d). For multi-view inspection, Dinomaly2 achieves 94.9% and 94.6% object-level AUROC on Real-IAD and MANTA-Tiny [22]. For inspection with 3D information (MVTec3D [23]) or infrared images (MulSen-AD [24]), Dinomaly2 achieves unprecedented 97.4% and 97.6% I-AUROC, which is higher than specialized multi-modal UAD methods. Beyond industrial domains, Dinomaly2 excels in medical images [25], microscopic images for life science [26], outdoor maintenance [27], and surveillance drone imagery [28]. Furthermore, Dinomaly2 demonstrates strong adaptability under few-shot constraints, presenting 98.7% and 97.4% I-AUROC on MVTec-AD and VisA using only 8 normal examples per class. Furthermore, we present the first systematic investigation of scalability in UAD, revealing that Dinomaly2 exhibits strong positive scaling behaviors across model size, resolution, and foundation quality (Fig. 1(e)).

A preliminary version of this work has been published in CVPR2025 (Dinomaly [29]). This work extends the initial version into a more comprehensive and powerful framework with the following major aspects: (1) Context-Aware Recentering, which resolves multi-class confusion through feature space transformation; (2) Natural extension to various modalities and settings, including multi-view, RGB-3D, RGB-IR, and few-shot UAD. achieving SOTA results without major modifications; (3) Introduction of the inference-unified MUAD setting, advancing toward complete unification where models operate across mixed categories with a single threshold; and (4) Comprehensive validation across 12 benchmark datasets spanning various modalities, settings, and domains, establishing Dinomaly2 as the a universal anomaly detection framework.

2 Related Work

Unsupervised Anomaly Detection for computer vision has evolved through several paradigms, each addressing the fundamental challenge of identifying abnormal patterns without explicit anomaly supervision. *Epistemic methods* are based on the assumption that the networks respond differently during inference between seen input and unseen input [20]. Within this paradigm, *pixel reconstruction* [30] methods assume that the networks trained on normal images can reconstruct anomaly-free regions well, but poorly for anomalous regions. However, such models may also incorrectly reconstruct anomalous regions if these regions resemble normal regions [8]. Therefore, *feature reconstruction (distillation)* [8], [31]–[33] is proposed to construct features of pre-trained encoders instead of raw pixels. Recent advances [14], [34] have explored diffusion models for pixel and feature reconstruction. *Synthetic anomaly* methods generate pre-defined defects on normal images to imitate anomalies, converting UAD into supervised classification [35] or segmentation tasks [36], [37]. For example, CutPaste [35] simulates anomalous regions by randomly pasting cropped patches of normal images, while DRAEM [36] and DesT-Seg [38] construct anomalous regions using Perlin noise as the mask and another image as the additive anomaly. SimpleNet [39] introduces anomaly by injecting Gaussian noise in the pre-trained feature space. These methods heavily rely on how well the synthetic anomalies match the real anomalies, which makes it hard to generalize to different datasets. *Feature statistics* (or memory-based) methods [7], [40], [41] memorize all normal features (or their modeled distribution) extracted by networks pre-trained on large-scale datasets and match them with test samples during inference, *e.g.*, PatchCore [7]. Since these methods require memorizing, processing, and matching nearly all features from training samples, they are computationally expensive in both training and inference, especially when the training set is large.

Multi-Class UAD was first introduced in UniAD [9], aiming to detect anomalies across different classes using a unified model. In this setting, conventional UAD methods often face the challenge of "identical shortcuts", where both anomaly-free and anomalous samples can be effectively recovered during inference [9]. Many current studies focus on addressing this challenge [9], [11]–[13]. UniAD [9] employs a neighbor-masked attention module and a feature-jitter strategy to mitigate these shortcuts. HVQ-Trans [12] proposes a vector quantization (VQ) Transformer model that induces large feature discrepancies for anomalies. LafitE [13] utilizes a latent diffusion model and introduces a feature editing strategy to alleviate this issue. DiAD [14] also employs diffusion models for multi-class UAD settings. ReContrast [11] alleviates the identity mapping by cross-reconstruction between two encoders. ViTAD [42] abstracts a unified feature-reconstruction UAD framework and employs Transformer building blocks. MambaAD [15] explores State Space Model (SSM) for multi-class UAD. Despite these efforts, existing unified models still exhibit a notable performance gap compared to SOTA single-class models, constraining the practicability of multi-class method.

Multi-Modal UAD. Building upon advances in 2D UAD, recent works have incorporated additional modalities beyond conventional 2D RGB data to enhance anomaly detection capabilities. Real-IAD [2] extends traditional paradigm to include multiple camera views for each object in real-world manufacturing environments. MVTec3D [23] explores 3D industrial inspection through point cloud data, addressing geometric anomalies that are invisible in 2D imagery. A number of methods were proposed to leverage both RGB and 3D information for UAD [16], [17], [43]. To further advance multi-modal UAD, Real-IAD D^3 [44] provides synchronized 2D, 3D, and pseudo-3D photometric stereo data. MulSen-AD [24] incorporates infrared (IR) images to exploit thermal information for anomaly detection.

Few-Shot UAD. Real-world scenarios often face data scarcity challenges, particularly during system cold-start or when dealing with rare object categories. Early few-shot UAD approaches leverage vision-language models to compensate for scarce visual data. For example, WinCLIP [18] utilizes CLIP's pre-trained vision-language representations and introduces window-based prompting strategies

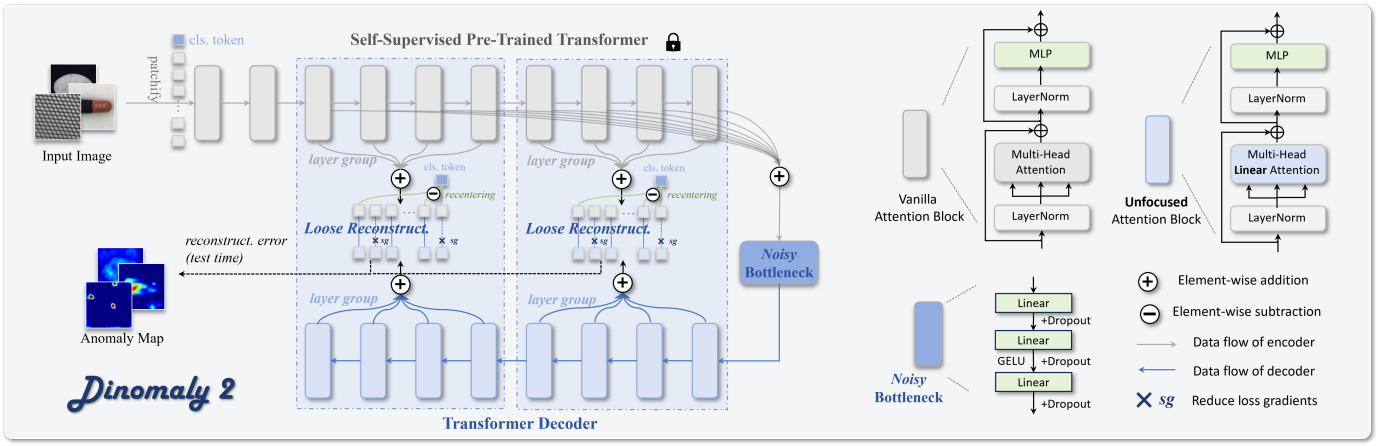

Fig. 2. The framework of Dinomaly2, built by simple and pure Transformer building blocks. A pre-trained encoder extracts multi-layer features, which are aggregated through a bottleneck and reconstructed by a decoder. During training, the model is optimized exclusively on normal images, while during inference, anomalies are detected based on reconstruction errors.

to adapt to specific object categories. PromptAD [45] extends this paradigm by designing learnable prompts that capture domain-specific patterns. Meta-learning approaches represent another direction. For example, InCTRL [19] and MetaUAS [46] employ meta-learning frameworks to learn transferable detection strategies from the auxiliary datasets and apply them to down-stream datasets. However, both multi-modal and few-shot extensions require specialized architectures and distinct pipelines, lacking a unified framework across diverse scenarios.

3 Method

3.1 Overview of Dinomaly2 Framework

The ability to recognize anomalies from what we know is an innate human capability, serving as a vital pathway for us to explore the world. Similarly, we construct a reconstruction-based framework that relies on the epistemic characteristic of artificial neural networks. Dinomaly2 consists of an encoder, a bottleneck, and a reconstruction decoder, as shown in Fig. 2.

Given an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$, a pre-trained Vision Transformer (ViT) $\mathcal{E}$ with L layers serves as the universal feature extractor:

$$\mathbf{F} = \mathcal{E}(\mathbf{x}) = \{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_L\}, \tag{1}$$

where $\mathbf{f}_i \in \mathbb{R}^{N\times d}$ represents the feature representation at the i-th layer, with N being the number of tokens and d the feature dimension. Without loss of generality, we take a ViT-Base [47] architecture with L=12 layers by default. We collect a group of features $\mathbf{F}^*$ from the L^*=8 middle layers $\mathcal{M} = \{3, ..., 10\}$ to capture multi-scale semantic information:

$$\mathbf{F}^* = \{\mathbf{f}_i \mid i \in \mathcal{M}\} = \{\mathbf{f}_3, \mathbf{f}_4, \dots, \mathbf{f}_{10}\}. \tag{2}$$

The bottleneck $\mathcal{B}$ is implemented as a simple multi-layer perceptron (MLP) that processes selected intermediate features:

$$\mathbf{z} = \mathcal{B}(\mathbf{z}_0) = \mathcal{B}(\sum_{i\in\mathcal{M}} \mathbf{f}_i), \tag{3}$$

where $\mathbf{z}$ is the output of $\mathcal{B}$. The decoder $\mathcal{D}$ consists of L_d Transformer layers that reconstruct the encoder's intermediate features:

$$\hat{\mathbf{F}} = \mathcal{D}(\mathbf{z}) = \{\hat{\mathbf{f}}_{10}, \hat{\mathbf{f}}_9, \dots, \hat{\mathbf{f}}_3\}, \tag{4}$$

where $\hat{\mathbf{f}}_i$ represents the reconstructed feature at layer i, with L_d=8 corresponding to the number of selected encoder layers.

During training, the model learns to reconstruct normal patterns by optimizing the feature-level reconstruction loss. The primary objective is to minimize the cosine distance between original and reconstructed features:

$$\mathcal{L}_{\text{recon}} = \frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} d_{\cos}(\mathcal{F}(\mathbf{f}_i), \mathcal{F}(\hat{\mathbf{f}}_i)), \tag{5}$$

where $\mathcal{F}$ denotes the flatten operation and $d_{\cos}$ denotes the cosine distance. The model is trained exclusively on normal samples $\mathbf{x}\sim\mathcal{X}_{\text{normal}}$, learning to capture the underlying distribution of normal patterns while remaining ignorant of anomalous variations. During inference, the decoder is expected to reconstruct normal regions of feature maps but fails for anomalous regions, as these patterns were never observed during training.

Recent studies [48], [49] have demonstrated that self-supervised Vision Transformers (ViTs) [50]–[52] provide more robust and universal features for UAD compared to domain-specific ImageNet features. In this paper, we present the first systematic investigation of the scaling behaviors of vision foundation models in UAD, as illustrated in Fig. 1(e). Our comprehensive evaluation encompasses pre-training strategy (Table 18), model size (Table 16), and input resolution (Table 17), which are detailed in the Experiment section.

Over-generalization. Prior studies [9]–[11] attribute the performance degradation of UAD methods trained on diverse multi-class samples to the "identity mapping" phenomenon. In this work, we reinterpret this issue as an "over-generalization" problem. Generalization ability is a desirable property of neural networks, allowing them to perform equally well on unseen test sets. However, excessive generalization is undesirable in the context of unsupervised anomaly detection that leverages the epistemic

nature of neural networks. In multi-class UAD settings, the increasing diversity of images and patterns may cause the decoder to over-generalize its reconstruction ability to unseen anomalous samples, thereby undermining anomaly detection based on reconstruction error.

3.2 Noisy Bottleneck

A direct solution for identity mapping is to shift "reconstruction" to "restoration". Specifically, instead of directly reconstructing the normal images or features given normal inputs, perturbations have been introduced as pseudo anomalies, either on input images [36], [38] or on features [9], [13]. The decoder is still required to restore anomaly-free images or features, formulating a denoising-like framework. However, such methods rely on heuristic and synthetic anomaly generation strategies that may not be generalize well across domains, datasets, and methods.

*"**Dropout is all you need.**"*

In this work, we turn to leveraging Dropout, a simple yet effective technique originally proposed to mitigate overfitting [53]. In Dinomaly2, Dropout is applied to randomly discard neural activations in an MLP bottleneck. The Noisy Bottleneck $\mathcal{B}$ is implemented as an MLP with L_b=3 layers. Given the aggregated features $\mathbf{z}_0 = \sum_{i\in\mathcal{M}} f_i$ from the encoder's layers, the $\mathcal{B}$ processes them through sequential transformations:

$$\mathbf{z} = (\mathbf{m}_{L_b}\odot\phi_{L_b})\circ(\mathbf{m}_{L_b-1}\odot\phi_{L_b-1})\circ\cdots\circ(\mathbf{m}_1\odot\phi_1)(\mathbf{z}_0), \quad (6)$$

where $\mathbf{m}\sim$ Bernoulli$(1-p)$ is a random binary mask, ϕ is a linear layer (with activation), and $\odot$ denotes element-wise multiplication.

The role of the Noisy Bottleneck can be explained as introducing *pseudo feature anomaly*. During training, Dropout creates an exponentially large ensemble of $2^{n\cdot L_d}$ subnetworks (where n is the number of units of each linear layer), each observing a different "corrupted" version of the features. This provides a parameter-free, domain-agnostic corruption mechanism, forcing the decoder to map corrupted inputs back to the most likely normal patterns.

The role of the Noisy Bottleneck can also be interpreted through the *information bottleneck* (IB) principle [54], [55]. The UAD task imposes dual constraints that distinguish between normal and anomalous input domains. The IB objective can be formulated as:

$$\max_{\mathcal{B},\mathcal{D}} I(\mathbf{z};\hat{\mathbf{F}}|\mathbf{x}\in\mathcal{X}_{\text{normal}}) - \beta I(\mathbf{z};\mathbf{F}^*|\mathbf{x}\in\mathcal{X}_{\text{all}}) \quad (7)$$

where $\mathcal{X}_{\text{normal}}$ denotes the normal domain, $\mathcal{X}_{\text{all}} = \mathcal{X}_{\text{normal}}\cup\mathcal{X}_{\text{ano}}$ denotes the complete input space, $I(\cdot;\cdot)$ denotes mutual information, and β controls the compression-reconstruction trade-off. While direct access to $\mathcal{X}_{\text{ano}}$ is unavailable, Dropout provides an implicit constraint on $I(\mathbf{z};\mathbf{F}^*|\mathbf{x}\in\mathcal{X}_{\text{ano}})$. The key insight is that Dropout does not uniformly reduce $\mathcal{B}$'s capacity; instead, it enforces a redundant encoding scheme specifically tailored to the statistical structure of $\mathcal{X}_{\text{normal}}$. Formally, the learned encoding maximizes:

$$\mathbb{E}_{\mathbf{M}}\left[I(\mathcal{B}(\mathbf{F}^*;\mathbf{M});\hat{\mathbf{F}}|\mathbf{x}\in\mathcal{X}_{\text{normal}})\right], \quad (8)$$

where $\mathbf{M} = \{\mathbf{m}_1,\ldots,\mathbf{m}_{L_b}\}$ denotes the collection of independent dropout masks. The expectation is taken over all possible realizations of dropout masks across all layers. Since the overall capacity of $\mathcal{B}$ is constrained, $I(\mathbf{z};\mathbf{F}^*|\mathbf{x}\in\mathcal{X}_{\text{ano}})$ is implicitly limited by the redundant encoding of observed normal patterns and the inefficient encoding of unseen anomalous patterns.

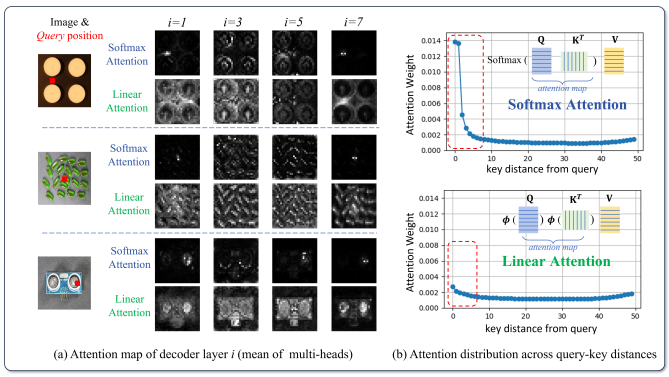

Fig. 3. Softmax Attention *vs.* Linear Attention. (a) Visualization of attention maps, and (b) Attention distribution, demonstrate both demonstrates Softmax Attention's tendency to focus on local regions versus Linear Attention's distributed attention pattern.

3.3 Unfocused Linear Attention

Softmax Attention [56] is the cornerstone of Transformer architectures, enabling models to dynamically capture relationships between different parts of the input. However, in the context of multi-class anomaly detection, the very strength of attention becomes a liability.

The Focusing Paradox. Formally, given an input sequence $\mathbf{f}\in\mathbb{R}^{N\times d}$ with length N, attention first transforms it into three matrices: the query matrix $\mathbf{Q}\in\mathbb{R}^{N\times d}$, the key matrix $\mathbf{K}\in\mathbb{R}^{N\times d}$, and the value matrix $\mathbf{V}\in\mathbb{R}^{N\times d}$:

$$\mathbf{Q} = \mathbf{f}\mathbf{W}^Q, \mathbf{K} = \mathbf{f}\mathbf{W}^K, \mathbf{V} = \mathbf{f}\mathbf{W}^V, \quad (9)$$

where $\mathbf{W}^Q,\mathbf{W}^K,\mathbf{W}^V\in\mathbb{R}^{d\times d}$ are learnable parameters. By computing the attention map by the query-key similarity, the output of Softmax Attention (SA) is given as: [1]

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Softmax}(\mathbf{Q}\mathbf{K}^T)\mathbf{V}. \quad (10)$$

Back to UAD, previous methods [9], [12] replace convolutions with attention, since convolutions are prone to learning identical mappings. Nevertheless, the softmax mechanism inherently encourages the model to concentrate on the most relevant features, forming identity mapping by only concentrating on corresponding input locations (Fig. 3).

*"**One man's poison is another man's meat.**"*

Linear Attention as a Feature, Not a Bug. Linear Attention (LA) was proposed as an alternative to reduce the computation complexity of vanilla Softmax Attention with respect to the number of tokens [57]. By replacing Softmax operation with a simple activation function $\phi(\cdot)$ (usually $\phi(x) = \text{elu}(x)+1$), the computation order changes from $(\mathbf{Q}\mathbf{K}^T)\mathbf{V}$ to $\mathbf{Q}(\mathbf{K}^T\mathbf{V})$. Formally, LA is given as:

$$\text{LA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = (\phi(\mathbf{Q})\phi(\mathbf{K}^T))\mathbf{V} = \phi(\mathbf{Q})(\phi(\mathbf{K}^T)\mathbf{V}), \quad (11)$$

1. The full form of Attention is Softmax$(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}$. The constant denominator and multi-head mechanism are omitted for narrative simplicity.

where the computation complexity is reduced from $\mathcal{O}(N^2d)$ to $\mathcal{O}(Nd^2)$. The trade-off between complexity and expressiveness is a dilemma. Previous studies [58], [59] attribute LA's performance degradation on supervised tasks to its incompetence in focusing on important regions related to the query, such as foreground or neighbors. This property, however, is exactly what the reconstruction decoder favors in our contexts.

In Dinomaly2, we leverage LA's inherent inability to focus as a desirable property for anomaly detection. To examine how how attentions propagate information, we train two variants of Dinomaly using vanilla SA or LA as the spatial mixer in the decoder and visualize their attention maps. As shown in Fig. 3(a), SA tends to focus on the exact region of the query, while LA spreads its attention across the entire image. From a frequency domain view, LA acts as an low-pass filter [60] that cannot selectively amplify local high-frequency details (Fig. 3(b)), compelling the network to reconstruct based on learned global patterns. This diffusive propagation prevents identity mapping in anomaly detection.

3.4 Context-Aware Recentering

A fundamental challenge in multi-class UAD arises from the context-dependent nature of anomalies: identical visual features can be normal in one context but anomalous in another. For example, a vehicle is normal on a highway but anomalous on a pedestrian walkway, whereas a pedestrian is expected on a sidewalk but becomes anomalous on a highway (Fig. 4(a)). This contextual ambiguity becomes particularly problematic in unified multi-class models, where the decoder may learn to reconstruct both vehicles and pedestrians in any context (Fig. 4(b)), undermining the anomaly detection capability.

*"**Beauty is in the eye of the beholder.**"*

To address this challenge, we introduce Context-Aware Recentering, a simple yet effective mechanism that conditions feature reconstruction on class-specific context. The key insight is to leverage the class token from Vision Transformers as a contextual anchor that encodes class-specific normal patterns. Formally, let $\mathbf{f}_i \in \mathbb{R}^{(N+1)\times d}$ denote the feature representation at encoder layer i, where the first token $\mathbf{f}_i^{cls} \in \mathbb{R}^d$ represents the class token and the remaining $\mathbf{f}_i^{patch} \in \mathbb{R}^{N\times d}$ represent patch tokens. Instead of reconstructing the patch features directly, we reconstruct the recentered features:

$$\tilde{\mathbf{f}}_i^{patch} = \mathbf{f}_i^{patch} - \mathbf{f}_i^{cls} \otimes \mathbf{1}_N \tag{12}$$

where $\otimes$ denotes the outer product and $\mathbf{1}_N$ is a vector of ones. This operation effectively shifts the origin of the feature space to be class-specific, with the class token serving as the new zero point (Fig. 4(c)). The decoder then learns to reconstruct these recentered features rather than the original ones.

Without any computation-overhead, this simple subtraction operation addresses the multi-class confusion problem through two mechanisms. First, by using class tokens as anchors, patch features from different classes are mapped to different reference frames that are unique in each scene, where the identical local pattern obtains different meanings.

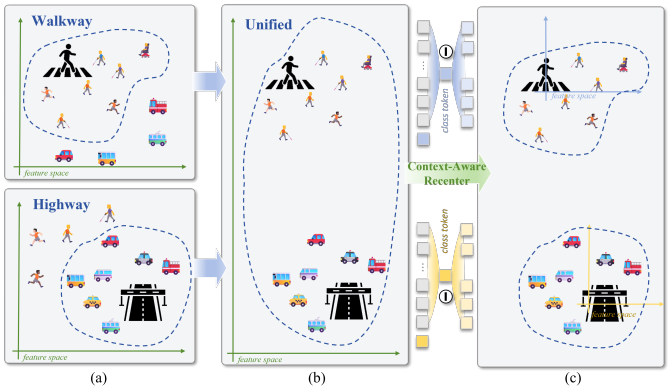

Fig. 4. Context-Aware Recentering and a toy example. (a) Separate detectors for sidewalk and highway show how identical objects (pedestrians/vehicles) may receive opposite anomaly labels depending on context. (b) A unified detector faces contextual ambiguity without proper conditioning. (c) Our Context-Aware Recentering uses class tokens as reference anchors to map patch features into class-specific coordinate systems, resolving multi-class confusion through simple subtraction.

Second, the subtraction operation implicitly conditions the reconstruction on class identity without requiring explicit class labels.

3.5 Loose Reconstruction

A critical challenge in multi-class UAD lies in the reconstruction objective design. Prior feature-reconstruction UAD methods [8], [33], [49] typically follow knowledge distillation paradigms [61], where decoder layers are trained to precisely mimic corresponding encoder layers. While this layer-to-layer supervision provides strong learning signals, it paradoxically enables the decoder to become too proficient at reconstruction, enabling it to restore even anomalous patterns that were never explicitly encountered during training. We propose Loose Reconstruction, which deliberately relaxes both structural constraints and optimization objectives to prevent over-generalization in multi-class settings.

*"**The tighter you squeeze, the less you have.**"*

Loose Constraint. Instead of enforcing strict layer-to-layer correspondence between encoder and decoder features (Fig. 5(a-c)), we group multiple layers into semantic clusters:

$$\mathbf{g}_k = \sum_{i\in\mathcal{S}_k} \mathbf{f}_i, k \in \mathcal{G} \tag{13}$$

where $\mathcal{S}_k$ denotes a group of layers. All layers may be grouped into a single cluster, so that $\mathcal{G} = \{1\}$, $\mathcal{S}_1 = \{3, 4, ..., 10\}$ (Fig. 5(d)). Given the hierarchical nature of Vision Transformers, where shallow layers capture low-level visual features and deeper layers encode high-level semantics, we aggregate features from multiple encoder layers: $\mathcal{G} = \{1, 2\}$, $\mathcal{S}_1 = \{3, 4, 5, 6\}$ and $\mathcal{S}_2 = \{7, 8, 9, 10\}$ represent shallow and deep layer groups respectively (Fig. 5(e)). The decoder then reconstructs these grouped representations rather than individual layers:

$$\hat{\mathbf{g}}_k = \sum_{j\in\mathcal{S}'_k} \hat{\mathbf{f}}_j, k \in \mathcal{G}. \tag{14}$$

This scheme loosens the layer-to-layer correspondence and grants the decoder with more degrees of freedom, so that

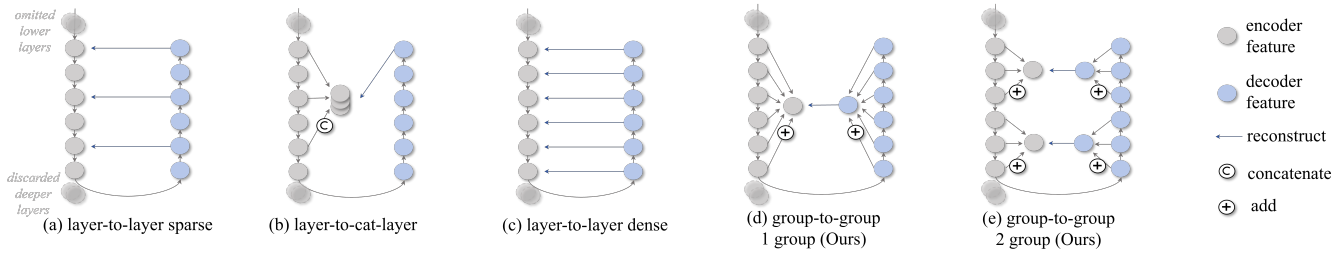

Fig. 5. Schemes of reconstruction constraint. (a) Layer-to-layer (sparse). (b) Layer-to-cat-layer. (c) Layer-to-layer (dense). (d) Loose group-to-group, 1-group (Ours). (e) Loose group-to-group, 2-group (Ours).

the decoder is allowed to act much more differently from the encoder when the input pattern is unseen.

Loose Loss. Beyond structural relaxation, we further introduce a selective optimization strategy that dynamically adjusts gradient flow based on reconstruction quality. Following [11], we adopt a hard-mining global cosine loss that selectively reduces gradients for well-reconstructed regions:

$$\mathcal{L}_{\text{loose}} = \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}} d_{\cos}(\mathcal{F}(\mathbf{g}_k), \mathcal{F}(\hat{\mathbf{g}}'_k)), \tag{15}$$

where $\hat{\mathbf{g}}'_k$ undergoes selective gradient modulation:

$$\hat{\mathbf{g}}'_k(n) = \begin{cases} \text{sg}_{0.1}(\hat{\mathbf{g}}_k(n)), & \text{if } d_{\cos}(\mathbf{g}_k(n), \hat{\mathbf{g}}_k(n)) > \tau_k \\ \hat{\mathbf{g}}_k(n), & \text{otherwise} \end{cases}, \tag{16}$$

where $\text{sg}_{0.1}(\cdot)$ reduces gradients to 10% of their original magnitude[2], and τ_k represents the k-th percentile threshold within each batch (90% by default). The combination of loose constraints and loose loss creates a reconstruction objective that is deliberately imperfect—capable of capturing normal patterns while maintaining sufficient reconstruction error on anomalous regions.

Anomaly Scoring. Finally, the reconstruction errors serve as the foundation for anomaly localization and detection. Given encoder feature groups $\mathbf{g}_k$ and their reconstructed counterparts $\hat{\mathbf{g}}_k$, we compute token-wise cosine similarity to generate anomaly maps:

$$\mathbf{A} = \frac{1}{|\mathcal{G}|}\sum_{k\in\mathcal{G}} \mathbf{A}_k, \tag{17}$$

$$\mathbf{A}_k(n) = 1 - \frac{\mathbf{g}_k(n)\cdot\hat{\mathbf{g}}_k(n)}{||\mathbf{g}_k(n)||_2 \cdot ||\hat{\mathbf{g}}_k(n)||_2}, \tag{18}$$

where n indexes the spatial tokens. For image-level anomaly detection, we compute the anomaly score as the mean of the top z% (default 1%) most anomalous pixels:

$$s_{image} = \text{mean}(\text{top}_{z\%}(\mathbf{A})). \tag{19}$$

3.6 Seamless Extension Beyond Plain 2D UAD

The minimalist philosophy of Dinomaly2 naturally extends beyond 2D images to diverse data modalities. Instead of introducing complex fusion mechanisms or modality-specific architectures, we demonstrate that minimal adaptations of our framework achieve state-of-the-art results.

*"**Entia non sunt multiplicanda praeter necessitatem.**"*

2. Complete gradient stopping occasionally causes training instability.

Multi-View Inspection. Industrial inspection often requires examining objects from multiple viewpoints. For multi-view datasets such as Real-IAD [2] and MANTA [22], each object is captured from V=5 camera angles. We keep the training paradigm untapped, including all views of all classes. During inference, anomaly maps from all views are concatenated, and the object-level anomaly score is computed as the mean of the top $z\% \cdot V$ most anomalous pixels across the concatenated anomaly map:

$$s_{object} = \text{mean}(\text{top}_{z\%\cdot V}(\bigcup_{v=1}^{V} \mathbf{A}_v)) \tag{20}$$

where $\mathbf{A}_v$ denotes the anomaly map from view v. This straightforward aggregation strategy consolidates multi-view information without requiring view-specific training or complex inter-view interaction.

Aligned Multi-Modality. For the datasets combining RGB images with pixel-aligned 3D point cloud data (*e.g.*, MVTec3D [23]), we adopt a minimal fusion strategy. The 3D point cloud is first rendered as a depth map using projection following [62]. Both RGB image and depth map are independently processed through the same pre-trained ViT encoder, yielding feature representations $\mathbf{F}^*_{rgb}$ and $\mathbf{F}^*_{depth}$ respectively. The multi-modal encoder representation is obtained through element-wise averaging:

$$\mathbf{F}^* = \frac{1}{2}(\mathbf{F}^*_{rgb} + \mathbf{F}^*_{depth}). \tag{21}$$

The subsequent bottleneck, decoder, and reconstruction processes remain identical to the standard framework. This naive fusion, requiring no additional parameters or architectural modifications, allows Dinomaly2 to leverage complementary geometric and appearance information.

Non-aligned Multi-Modality. In some scenarios, multi-modal information is acquired without spatial alignment, where images of different modalities are acquired from different camera angles. Here, we adopt a straightforward extension of our multi-view inspection strategy: each modality is treated as an independent view. For example, in MulSen-AD [24], RGB and infrared (IR) images are processed as separate views within our unified framework. The final object-level anomaly score is computed as the sum of the individual image-level anomaly scores from both modalities.

Few-Shot Anomaly Detection. Real-world UAD scenarios often encounter data scarcity, where only limited normal samples are available. Dinomaly2's minimalist design naturally adapts to multi-class few-shot setting without any architectural modifications. Our adaptation involves only

the simplest enhancement: standard data augmentation techniques during training, including random flipping, rotation, and translation. These elementary augmentations expand the limited normal sample space without introducing domain-specific assumptions (text-prompts) [18] or complex meta-learning strategies [46] employed by specialized few-shot UAD methods.

4 Experiments

4.1 Experimental Settings

We conduct a comprehensive evaluation across a wide range of benchmarks, referred to as our UAD Dataset Marathon (Table 1). Specificially, **MVTec-AD** [1] is the most widely-used industrial anomaly detection benchmark, containing 5 texture classes and 10 object classes. **VisA** [21] focuses on complex industrial products with challenging anomaly patterns, *e.g.*, multiple objects in one image. **MPDD** [63] targets visual defect detection in metal parts manufacturing with real-world industrial conditions. **BTAD** [64] contains real-world industrial products with inherently noisy training sets that include some anomalous samples [65].

Real-IAD [2] is a large multi-view dataset containing 30 categories, each with 5 camera views. We follow the official splitting. **MANTA** [22] is a large-scale multi-view dataset for tiny objects (e.g. various seeds and beans). We use the official MANTA-Tiny split, which is more accessible. **MVTec3D** [23] contains point clouds scanned by industrial 3D sensors with paired 2D RGB images from 10 categories. 3D point clouds are pre-processed to depth maps following [62]. **MulSen-AD** [24] is a multi-sensor dataset unifying data from RGB cameras, laser scanners, and lock-in infrared thermography. We utilize RGB and IR images.

Uni-Medical [66] is a medical UAD dataset consisting of 2D slices of brain CT, liver CT, and retinal OCT, extracted from [25]. **ApoCell** consists of confocal fluorescence microscopy images from *C. elegans* embryos, derived from [26]. Normal images contain only viable cells, while anomalous images include cells undergoing apoptosis. **MIAD** [27] consists of outdoor images with uncontrolled factors (viewpoint, background, etc.), focusing on maintenance inspection to keep equipment in optimum working condition. **Drone-Anomaly** [28] is a drone-based dataset for abnormal event surveillance captured by UAV cameras. This dataset contains 7 distinct scenes including highway, crossroad, bike roundabout, etc.

Metrics. Following prior works [15], [49], we adopt seven evaluation metrics. Image-level (I-) anomaly detection performance is measured by the Area Under the Receiver Operator Curve (AUROC), Average Precision (AP), and F_1 score under optimal threshold (F_1-max). Pixel-level (P-) anomaly localization is measured by AUROC, AP, F_1-max and the Area Under the Per-Region-Overlap (AUPRO). The results of a dataset is the average of all categories.

Implementation Details. By default, we used ViT-S/B/L pre-trained with DINOv2-registers (DINOv2-R) [67] as the encoder. The middle 8 layers of 12-layer ViT-B and ViT-S are used for reconstruction and feeding the bottleneck. ViT-Large contains 24 layers; therefore, we use every other two layers: $\mathcal{M} = \{5, 7, 9, ...19\}$. The decoder always contains 8 layer. Unless otherwise specified, a single model is trained jointly for all categories in each dataset.

TABLE 1
Statistics of 12 UAD datasets. I-ROC (*i.e.*, I-AUROC): the image-level AUROC of Dinomaly2 for preview.

Dataset	Modality	#Class	#Train/Test(good/bad)	Train Iters	I-ROC
MVTec-AD [1]	RGB	15	3,629/467/1,258	4×10^4	99.9
VisA [21]	RGB	12	8,659/962/1,200	4×10^4	99.3
MPDD [63]	RGB	6	888/176/282	2×10^4	99.0
BTAD [64]	RGB	3	1,799/451/290	2×10^4	96.3
Real-IAD [2]	multi-view	30	36,465/63,256/51,329	1×10^5	92.1
MANTA-Tiny [22]	multi-view	38	21,483/33,560/11,025	1×10^5	93.5
MVTec3D [23]	RGB-3D	10	2,656/249/948	4×10^4	97.4
MulSen-AD [24]	RGB-IR	15	1,391/153/491	2×10^4	97.6
Uni-Medical [66]	Gray	3	13,339/2,514/4,499	4×10^4	86.7
ApoCell [26]	Gray	1	2,704/1,597/1,016	1×10^4	83.2
MIAD [27]	RGB	7	70,000/17,500/17,500	1×10^5	84.5
Drone-Anomaly [28]	RGB	7	51,635/19,724/16,119	1×10^4	84.8

StableAdamW optimizer [68] is utilized with lr (learning rate)=2e-3, β=(0.9,0.999), wd (weight decay)=1e-4 and eps=1e-10. Under the MUAD setting, the model is trained with the batch size of 16 for It iterations as in Table 1. The epochs can be derived by $\frac{16 \cdot It}{\#\text{Train}}$. The lr warms up from 0 to 2e-3 in the first 100 iterations and cosine anneals to 2e-4 throughout the training. In the Noisy Bottleneck, the dropout rate is set to 0.2 or 0.4 based on simple parameter searching. Loose constraint with 2 groups is used by default. The discarding rate controling τ_k linearly increases from 0% to 90% in the first 1,000 iterations as warm-up. The input image size is 392^2 by default, except 280^2 for MANTA, Uni-Medical, and ApoCell with smaller original images. The mean of the top 1% pixels in an anomaly map is used as the image-level anomaly score. For Real-IAD with tiny anomalies, we use top 0.1%. Codes are implemented with Python 3.8 and PyTorch 1.12.0 cuda 11.3, and experiments were run on NVIDIA GeForce RTX3090/4090 GPUs (24GB). Most results of compared MUAD results are cited from a benchmark paper ADer [66], to which we express our gratitude. Code will be available at: https://github.com/guojiajeremy/Dinomaly.

4.2 Comparison Results

(1) Multi-Class UAD on 2D Industrial Datasets. We first evaluate Dinomaly2 against state-of-the-art multi-class UAD methods on conventional 2D industrial datasets. As shown in Table 2, Dinomaly2-B (equipped with ViT-B) already exceeds previous methods by a large margin. Our strongest model Dinomaly2-L (equipped with ViT-L) achieves exceptional performance across all benchmarks, with I-AUROC reaching 99.9%, 99.3%, 99.0% and 96.0% on MVTec-AD [1], VisA [21], MPDD [63], and BTAD [64], respectively.

On MVTec-AD, our unified multi-class model achieves near-perfect detection performance. Dinomaly2-L surpasses previous SOTA method, MambaAD [15] by a large margin (99.9% vs. 98.6%) in I-AUROC and significantly outperforms in pixel-level metrics, achieving 70.3% AP and 95.1% AUPRO. On VisA, Dinomaly2-L delivers substantial improvements over the previous best method ReContrast [11], with gains of 3.8% in I-AUROC and 3.5% in AUPRO. The remarkable performance on MVTec-AD and VisA suggests that these datasets may be approaching saturation under our approach. The performance gains are also particularly notable on the MPDD dataset, where our method achieves

TABLE 2
Multi-class UAD performance on conventional 2D industrial datasets (%). † denotes methods originally designed for single-class UAD but trained under MUAD settings. Underlines represent runner-up performance excluding Dinomaly variants.

Method	**MVTec-AD** [1] (15 classes)							**VisA** [21] (12 classes)						
	Image-level			Pixel-level				Image-level			Pixel-level			
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
RD4AD† [8]	94.6	96.5	95.2	96.1	48.6	53.8	91.1	92.4	92.4	89.6	98.1	38.0	42.6	91.8
SimpleNet† [39]	95.3	98.4	95.8	96.9	45.9	49.7	86.5	87.2	87.0	81.8	96.8	34.7	37.8	81.4
DeSTSeg† [38]	89.2	95.5	91.6	93.1	54.3	50.9	64.8	88.9	89.0	85.2	96.1	39.6	43.4	67.4
UniAD [9]	96.5	98.8	96.2	96.8	43.4	49.5	90.7	88.8	90.8	85.8	98.3	33.7	39.0	85.5
ReContrast [11]	98.3	99.4	97.6	97.1	60.2	61.5	93.2	95.5	96.4	92.0	98.5	47.9	50.6	91.9
DiAD [14]	97.2	99.0	96.5	96.8	52.6	55.5	90.7	86.8	88.3	85.1	96.0	26.1	33.0	75.2
ViTAD [49]	98.3	99.4	97.3	97.7	55.3	58.7	91.4	90.5	91.7	86.3	98.2	36.6	41.1	85.1
MambaAD [15]	98.6	99.6	97.8	97.7	56.3	59.2	93.1	94.3	94.5	89.4	98.5	39.4	44.0	91.0
Dinomaly [29]	99.6	99.8	99.0	98.4	69.3	69.2	94.8	98.7	98.9	96.2	98.7	53.2	55.7	94.5
Dinomaly2-B	99.8	99.9	99.3	98.4	69.3	68.9	94.8	99.2	99.3	97.0	99.1	52.9	56.4	**95.5**
Dinomaly2-L	**99.9**	**100.0**	**99.5**	**98.6**	**70.3**	**69.9**	**95.1**	**99.3**	**99.4**	**97.2**	**99.2**	**54.7**	**57.1**	95.4

Method	**MPDD** [63] (6 classes)							**BTAD** [64] (3 classes)						
	Image-level			Pixel-level				Image-level			Pixel-level			
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
RD4AD† [8]	90.3	92.8	90.5	98.3	39.6	40.6	95.2	94.1	96.8	93.8	98.0	57.1	58.0	**79.9**
SimpleNet† [39]	90.6	94.1	89.7	97.1	33.6	35.7	90.0	94.0	97.9	93.9	96.2	41.0	43.7	69.6
DeSTSeg† [38]	92.6	91.8	92.8	90.8	30.6	32.9	78.3	93.5	96.7	93.8	94.8	39.1	38.5	72.9
UniAD [9]	80.1	83.2	85.1	95.4	19.0	25.6	83.8	94.5	98.4	94.9	97.4	52.4	55.5	78.9
ReContrast [11]	90.9	94.5	91.0	98.8	46.8	48.9	95.6	94.8	98.9	95.5	97.2	56.8	56.2	76.4
DiAD [14]	85.8	89.2	86.5	91.4	15.3	19.2	66.1	90.2	88.3	92.6	91.9	20.5	27.0	70.3
ViTAD [49]	87.4	90.8	87.0	97.8	44.1	46.4	95.3	94.0	97.0	93.7	97.6	58.3	56.5	72.8
MambaAD [15]	89.2	93.1	90.3	97.7	33.5	38.6	92.8	92.9	96.2	93.0	97.6	51.2	55.1	77.3
Dinomaly [29]	97.2	98.4	96.0	99.1	59.5	59.4	96.6	95.4	98.4	95.6	97.8	70.1	68.0	76.5
Dinomaly2-B	97.6	98.6	96.6	99.2	63.5	62.6	96.8	97.4	98.9	97.0	98.1	73.5	**68.4**	**79.9**
Dinomaly2-L	**99.0**	**99.4**	**97.8**	**99.3**	**64.6**	**62.8**	**97.1**	**97.5**	**99.2**	**97.4**	**98.2**	**73.8**	68.3	79.8

TABLE 3
Multi-view multi-class UAD performance on Real-IAD and MANTA-Tiny (%).

Method	**Real-IAD** [2] (30 classes)								**MANTA-Tiny** [22] (38 classes)							
	Object	Image-level			Pixel-level				Object	Image-level			Pixel-level			
	AUROC	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO	AUROC	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
RD4AD [8]	85.7	83.0	79.6	74.4	97.3	25.8	33.5	90.4	78.1	80.5	59.7	62.1	91.9	28.5	33.6	76.6
SimpleNet [39]	61.2	57.2	53.4	61.5	75.7	2.8	6.5	39.0	59.6	58.3	34.8	43.3	66.3	5.8	9.3	30.3
DeSTSeg [38]	89.0	82.3	79.2	73.2	94.6	37.9	41.7	40.6	66.7	65.3	40.9	46.8	71.2	9.8	8.7	27.0
UniAD [9]	87.3	83.0	80.9	74.3	97.3	21.1	29.2	86.7	78.2	80.5	59.7	62.5	89.8	22.6	29.5	76.3
ViTAD [49]	88.8	82.7	80.2	73.7	97.2	24.3	32.3	84.8	88.9	88.1	75.8	71.6	94.9	40.1	43.9	80.6
MambaAD [15]	89.9	86.3	84.6	77.0	98.5	33.0	38.7	90.5	86.3	86.1	73.1	69.6	94.1	37.0	41.4	80.7
MVAD [69]	90.2	86.6	84.8	77.2	97.9	30.3	36.8	91.2	84.4	84.6	69.4	67.3	93.5	32.5	38.2	80.3
Dinomaly [29]	92.2	89.3	86.8	80.2	98.8	42.8	47.1	93.9	92.1	91.1	82.1	77.0	95.1	44.8	48.5	84.1
Dinomaly2-B	94.3	91.6	89.8	82.6	**99.2**	47.8	51.1	96.0	93.2	92.2	83.7	79.0	96.1	51.6	53.1	88.3
Dinomaly2-L	**94.9**	**92.1**	**90.2**	**83.4**	**99.2**	**48.4**	**51.8**	**96.2**	**94.6**	**93.5**	**86.0**	**81.0**	**96.7**	**54.1**	**55.5**	**89.3**

99.0% I-AUROC, representing a 6.4% improvement over the previous SOTA method DeSTSeg [38] (92.6%).

(2) Multi-View Multi-Class UAD. Real-world industrial inspection often requires examining objects from multiple viewpoints to ensure comprehensive defect detection. On the Real-IAD [2] and MANTA-Tiny [22], Dinomaly2 demonstrates superior performance across all evaluation levels. As presented in Table 3, our method achieves 94.9% object-level AUROC on Real-IAD and 94.6% object-level AUROC on MANTA-Tiny. These results substantially outperform previous methods, including the MVAD [69] with multi-view interaction operation designed specifically for multi-view inspection.

(3) RGB-3D Multi-Class UAD. Extending beyond 2D imagery, we evaluate Dinomaly2 on MVTec3D [23], which combines RGB images with 3D point cloud data for industrial inspection. As shown in Table 4, our RGB-only model already achieves 95.0% I-AUROC, which not only surpasses previous RGB-based methods by a significant margin, but also reaches competitive performance compared to SOTA methods that leverage RGB+3D input (94.4%).

More importantly, our simple RGB-3D fusion strategy, *i.e.*, element-wise averaging of features from independently processed RGB images and depth maps, achieves unprecedented results. Dinomaly2-L attains 97.4% I-AUROC and 98.6% AUPRO, outperforming previous multi-modal-specific multi-class SOTA GLFM (94.4%, 93.1%) by a large margin. This remarkable performance validates our hypothesis that the minimalist philosophy naturally scales across data modalities without requiring specialized architectural modifications.

(4) RGB-IR Multi-Class UAD. For industrial scenarios requiring thermal information, we evaluate Dinomaly2 on MulSen-AD. As shown in Table 5, our straightforward approach achieves exceptional performance. Dinomaly2-B attains 97.6% object-level AUROC, substantially outperforming the specialized multi-sensor method TripleAD (94.2%).

(5) Single-Class UAD. To demonstrate that unified models do not compromise performance compared to specialized single-class approaches, we evaluate Dinomaly2 under conventional single-class settings. As reported in Table 6, our multi-class trained model achieves performance on

TABLE 4
Multi-class UAD performance for **RGB-3D** inspection on MVTec3D [23] with 10 classes (%). Results are reported for both RGB-only and RGB–3D Dinomaly2. † denotes method specifically designed for RGB+3D UAD.

Modality	Method	I-ROC	P-ROC	P-PRO
RGB	RD4AD [8]	79.3	98.3	93.5
	DeSTSeg [38]	82.5	82.1	65.1
	UniAD [9]	77.0	96.8	89.4
	ViTAD [49]	79.7	98.0	91.3
	MambaAD [15]	85.8	98.6	94.1
	Dinomaly2-B	93.9	**99.2**	**97.4**
	Dinomaly2-L	**95.0**	99.1	97.0
RGB+3D	BTF† [16]	68.2	96.8	90.4
	M3DM† [17]	72.4	96.0	87.2
	Shape-G† [70]	89.8	97.5	92.5
	CPMF† [71]	91.8	96.8	90.4
	GLFM† [72]	94.4	97.9	93.1
	Dinomaly2-B	97.0	**99.6**	**98.6**
	Dinomaly2-L	**97.4**	**99.6**	**98.6**

TABLE 5
Multi-class UAD performance for **RGB-IR** inspection on MulSen-AD [24] with 15 classes (%). † denotes method specifically designed for multi-sensor UAD. O-ROC: object-level AUROC. Pixel-level metrics are calculated respectively on RGB/Infra-red image.

Modality	Multi-class Unified	Method	O-ROC	P-ROC	P-AP
RGB	✗	PatchCore [7]	83.7	-	-
	✗	RD++ [8]	86.2	-	-
	✗	SimpleNet [39]	85.3	-	-
RGB+IR	✗	TripleAD† [24]	94.8	98.2/97.2	34.3/32.2
	✓	RD4AD [8]	85.0	98.5/98.3	27.4/28.2
	✓	ViTAD [49]	90.5	97.6/96.3	22.6/21.2
	✓	TripleAD† [24]	94.2	98.3/97.9	37.2/32.6
	✓	**Dinomaly2-B**	**97.6**	**99.2**/98.7	**39.8**/41.1
	✓	**Dinomaly2-L**	97.0	**99.2**/**98.8**	39.3/**41.8**

par with, or even superior to dedicated single-class methods. This result demonstrates that Dinomaly2 successfully bridges the long-standing performance gap between multi-class and single-class UAD methods, establishing unified models practically viable for real-world deployment.

(6) Few-Shot Multi-Class UAD. In resource-constrained scenarios where only limited normal samples are available, Dinomaly2 demonstrates exceptional few-shot learning capabilities. Using only 4 normal samples per class, our method achieves 98.1% I-AUROC on MVTec-AD and 96.7% on VisA, as shown in Table 7. These results even outperform the method that was specially designed for multi-class few-shot anomaly detection (IIPAD [73]: 96.1%, 88.3%). Moreover, our approach seamlessly integrates few-shot UAD with RGB-3D data. On the MVTec3D dataset, our method achieves an I-AUROC of 90.3% using only 4 samples per class, substantially outperforming the current SOTA [74] (74.0%)

(7) Inference-Unified Multi-Class UAD. Existing multi-class evaluation protocols often employ class-separate inference and evaluation. Here, we introduce the inference-unified setting where models must detect anomalies across mixed categories using the same threshold as if there is only one scene. The crux of UAD models lies in the mismatch of anomaly score distributions across different object categories. Although the normal and anomalous samples are well separated within each category, they become inseparable in the unified score distribution, as shown in Fig. 6.

TABLE 6
Single-class UAD performance (%).Multi-class Dinomaly2 results are also presented for comparison.

	Multi-class Unified	Method	I-ROC	P-ROC	P-PRO
MVTec-AD	✗	RD4AD [8]	98.5	97.8	93.9
	✗	PatchCore [7]	99.1	98.1	93.5
	✗	SimpleNet [32]	99.6	98.1	90.0
	✗	**Dinomaly2-B**	**99.8**	**98.4**	**95.3**
	✓	Dinomaly2-B	*99.8*	*98.4*	*95.1*
VisA	✗	RD4AD [8]	96.0	90.1	70.9
	✗	GLASS [37]	98.8	98.8	92.8
	✗	DiffusionAD [34]	98.8	98.9	96.0
	✗	**Dinomaly2-B**	**99.2**	**99.2**	**95.7**
	✓	Dinomaly2-B	*99.2*	*99.1*	*95.5*

TABLE 7
Few-shot multi-class UAD performance (%). The multi-class unified model is built using K-shot per class. Results of Dinomaly2 are the average over five experiments with random selected normal shots. † denotes method specifically designed for few-shot UAD.

	Shot/cls	Method	I-ROC	P-ROC	P-PRO
MVTec-AD	4	PatchCore [7]	74.9	92.6	80.8
		WinCLIP† [18]	94.0	92.9	84.4
		PromptAD† [45]	90.6	92.4	84.6
		InCTRL† [19]	94.5	-	-
		IIPAD† [73]†	96.1	97.0	91.2
		Dinomaly2-B	**98.1**	**97.4**	**93.5**
	8	**Dinomaly2-B**	98.7	97.5	93.6
	16	**Dinomaly2-B**	99.0	97.6	93.7
	Full	MambaAD [15]	98.6	97.7	93.1
VisA	4	PatchCore [7]	62.6	85.4	70.6
		WinCLIP† [18]	86.1	95.2	82.1
		PromptAD† [45]	88.8	97.2	84.7
		InCTRL† [19]	90.2	-	-
		IIPAD† [73]	88.3	97.4	88.3
		Dinomaly2-B	**96.7**	**98.2**	**94.4**
	8	**Dinomaly2-B**	97.4	98.4	94.8
	16	**Dinomaly2-B**	98.0	98.4	94.8
	Full	ReContrast [11]	95.5	98.5	91.9
MVTec3D	4	BTF [16]	64.3	97.6	90.4
		Shape-G [70]	69.8	97.9	91.8
		AST [62]	53.5	91.4	73.9
		CLIP3D-AD† [74]	74.0	97.0	90.2
		Dinomaly2-B	**90.3**	**99.3**	**97.6**
	8	**Dinomaly2-B**	93.4	99.5	98.1
	16	**Dinomaly2-B**	94.2	99.5	98.3
	Full	GLFM [72]	94.4	97.9	93.1

As shown in Table 8, Dinomaly2-B achieves 98.9% I-AUROC on MVTec-AD and 97.8% on VisA under this challenging setting, significantly outperforming previous methods. This evaluation better reflects real-world deployment scenarios where models encounter mixed object categories without prior knowledge of class identity.

(8) Comparison with Dinomaly-Based Extensions. Our framework is not only easy to use but also highly extensible. Since the release of our Dinomaly codebase, numerous subsequent methods have built upon the framework. As shown in Table 9, we compare Dinomaly2 with these recent techniques and extensions, including intrinsic normal prototype (INP) [75], post training cost-filtering mechanisms [76], and inter-view attention modules [77]. While these extensions provide modest improvements over the original Dinomaly, Dinomaly2 consistently outperforms all variants. This comparison validates that the simple yet principled improvements in Dinomaly2, particularly Context-Aware Recentering and optimized training hyperparameters, pro-

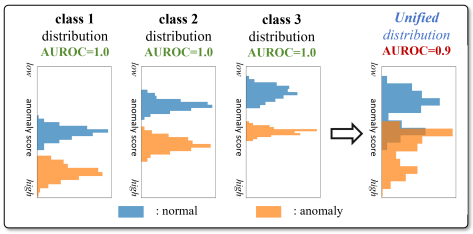

Fig. 6. The challenge of inference-unified multi-class UAD.

TABLE 8
Inference-unified multi-class UAD performance (%). The unified model is evaluated with all images in a dataset mixed together. Metrics represent performance across the entire mixed dataset rather than category-wise averages.

Method	MVTec-AD (Mixed)			VisA (Mixed)		
	I-ROC	P-ROC	P-AP	I-ROC	P-ROC	P-AP
RD4AD [8]	85.0	94.5	49.0	83.9	94.9	33.5
DRAEM [36]	82.3	81.8	41.1	70.0	82.7	27.0
PatchCore [7]	89.0	84.4	25.6	84.1	94.0	33.1
UniAD [32]	91.1	96.4	51.1	89.0	97.9	34.7
ReContrast [11]	90.7	96.8	56.3	93.2	97.1	42.1
ViTAD [49]	88.9	95.4	52.2	83.3	94.7	37.2
Dinomaly2-B	**98.9**	**97.3**	**62.4**	**97.8**	**98.7**	**48.6**

TABLE 9
Comparison with new SOTAs based on Dinomaly (%). These methods are developed or integrated on the top of our Dinomaly [29]. Results are presented as I-AUROC/AUPRO.

Method	MVTec-AD	VisA	Real-IAD
Dinomaly [29]	99.6/94.8	98.7/94.5	89.3/93.9
+INP [75]	99.7/94.9	98.9/94.4	90.5/95.0
+CostFilter [76]	99.7/94.8	98.7/94.7	-
+Inter-View Attention [77]	-	-	89.7/94.4
Dinomaly2-B	**99.8/95.1**	**99.2/95.5**	**91.1/96.0**
Dinomaly2-L	**99.9/95.1**	**99.3/95.4**	**92.1/96.2**

vide more substantial performance gains than complicated modifications and designs.

(9) Beyond Industrial Domain. We further evaluate Dinomaly2's universality across diverse application domains, as shown in Table 10. On the medical imaging dataset Uni-Medical [66], our method achieves SOTA performance for detecting pathological conditions in brain CT, liver CT, and retinal OCT scans. In microscopic images for biological research [26], Dinomaly2 attains 83.2% I-AUROC in identifying apoptotic cells in *C. elegans* embryos. In outdoor maintenance scenarios on MIAD [27] with uncontrolled environmental factors, our method reaches 84.5% I-AUROC. For aerial surveillance on the Drone-Anomaly dataset [28], Dinomaly2 successfully identifies abnormal events in UAV imagery across 7 distinct scenes. These cross-domain results demonstrate that the minimalist design philosophy and universal feature representations enable effective anomaly detection across vastly different data distributions and application contexts, positioning Dinomaly2 as a universal solution for real-world anomaly detection applications.

(10) Qualitative Visualization. We visualize the output anomaly maps of our Dinomaly2, as shown in Fig. 7. Anomaly maps are normalized to [0, 1] on a per-image basis using mean-max normalization.. It is noted that all our visualized samples are randomly chosen without cherry picking.

TABLE 10
Multi-class UAD performance on Uni-Medical [66], ApoCell [26], MIAD [27], and Drone-Anomaly [28] (%).

Method	I-ROC	P-ROC	P-PRO
Uni-Medical			
RD4AD [8]	76.4	96.4	86.5
DeSTSeg [38]	82.0	83.2	66.1
UniAD [9]	80.4	96.5	85.8
DiAD [14]	82.9	96.0	85.4
ViTAD [49]	81.5	97.0	86.5
Dinomaly2-B	**86.7**	**97.9**	**90.0**
MIAD			
RD4AD [8]	64.2	84.1	69.3
DeSTSeg [38]	68.7	72.6	16.7
UniAD [9]	61.2	91.1	69.0
ReContrast [11]	68.9	86.7	72.9
ViTAD [49]	71.5	80.9	60.4
Dinomaly2-B	**84.5**	**92.1**	**81.5**

Method	I-ROC	I-AP	I-F1
Drone-Anomaly			
ANDT† [28]	68.6	-	-
MADViT† [78]	74.2	-	-
RD4AD [8]	68.3	36.9	48.3
UniAD [9]	70.4	44.5	53.8
ViTAD [49]	77.0	56.4	59.4
Dinomaly2-B	**84.8**	**67.0**	**73.6**
ApoCell			
RD4AD [8]	80.4	70.3	68.7
DeSTSeg [38]	48.1	36.3	56.7
UniAD [9]	47.3	35.2	56.1
ReContrast [11]	81.6	70.3	71.1
ViTAD [49]	75.1	63.3	65.1
Dinomaly2-B	**83.2**	**71.3**	**73.8**

4.3 Ablation Study

(1) Overall Component Analysis. We systematically evaluate the effectiveness of each proposed component through comprehensive ablation studies on MVTec-AD and VisA datasets. As shown in Table 11, our baseline achieves competitive performance of 98.73% I-AUROC on MVTec-AD and 96.23% on VisA, demonstrating the strong foundation provided by our simple architecture. Each component contributes meaningfully to the overall performance. The Noisy Bottleneck (NB) provides a foundational gain, validating our hypothesis that simple Dropout-based noise injection effectively prevents over-generalization in multi-class settings. Context-Aware Recentering (CR) shows strong improvements on VisA (+1.19% I-AUROC), where the multi-class confusion problem is more pronounced. Unfocused Linear Attention (UA) and Loose Constraint (LC) demonstrate synergistic effects with other components. While UA and LC alone provides modest gains, their combination with NB and other components yields substantial improvements. Loose Loss (LL) consistently improves performance across both datasets, which validates our selective optimization strategy that reduces gradient flow for well-reconstructed regions. The full combination of all components achieves the best performance: 99.74%/99.82%/99.30% on MVTec-AD and 98.89%/99.02%/96.47% on VisA with only 10,000 training iterations, demonstrating that Dinomaly2 effectively addresses the challenges in multi-class anomaly detection through simple yet effective designs.

(2) Ablation on Other Architectures. To demonstrate the generalizability of our proposed components beyond the Dinomaly framework, we systematically integrate each module into ViTAD [49], a representative Transformer-based UAD method. As shown in Table 12, each Dinomaly2 component provides consistent improvements. The progressive integration yields substantial gains: starting from 98.27% to 99.25% I-AUROC—a total improvement of 0.98% I-AUROC and 2.03% AUPRO. This transferability analysis validates that our design principles are not framework-specific optimizations but applicable to diverse UAD architectures.

(3) Individual Component Analysis. We conduct ablations for Noisy Bottleneck as presented in Table 13. Experimental results demonstrate that Dinomaly2 is highly robust to different levels of dropout rate, performing optimally with dropout rates between 0.2-0.4. We also evaluate the effectiveness of other kind of noises, such as synthetic image anomaly (DRAEM-like [36]), feature jitter [9] (Gaussian

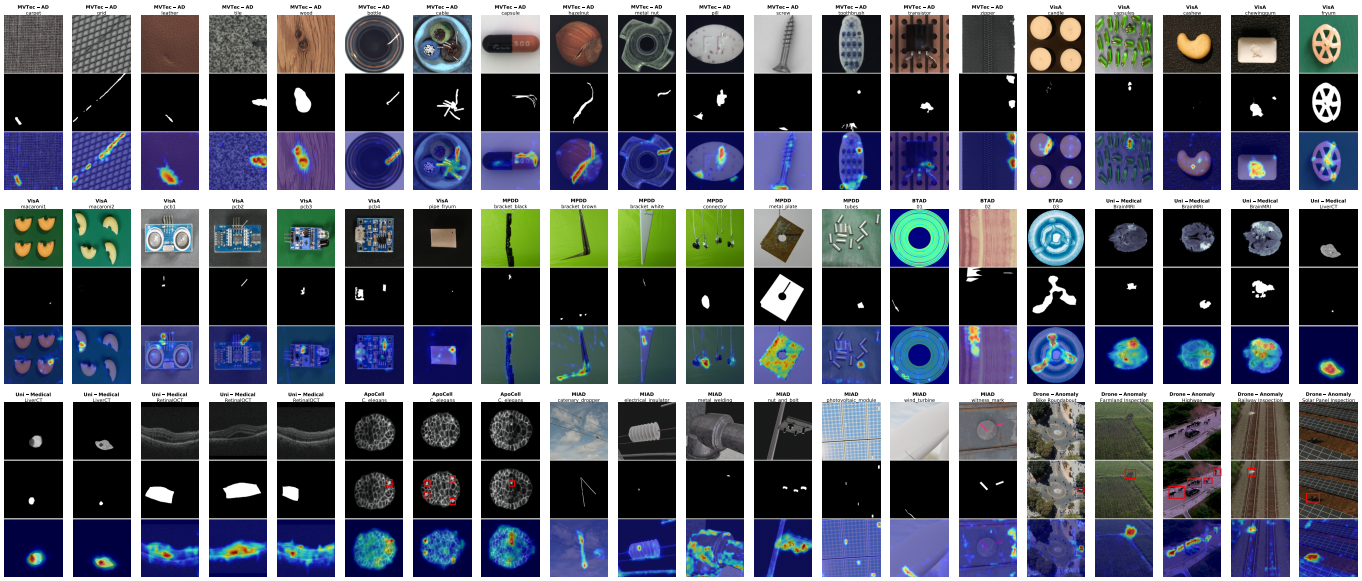

Fig. 7. Qualitative anomaly localization results across diverse domains. Anomaly maps are mean-max normalized to [0,1] range with red indicating high anomaly likelihood. All samples are randomly selected **without cherry-picking**.

TABLE 11
Systematic ablation study of proposed components on MVTec-AD and VisA (%). NB: Noisy Bottleneck. UA: Unfocused Linear Attention. LC: Loose Constraint (2 groups). LL: Loose Loss. CR: Context-Aware Recentering. ViT-B is used as the backbone. Training is reduced to 10,000 iterations for efficiency. Image-level and pixel-level metrics are reported as AUROC/AP/F_1-max and AUROC/AP/F_1-max/AUPRO, respectively.

NB	UA	LC	LL	CR	MVTec-AD		VisA	
					Image-Level	*Pixel-Level*	*Image-Level*	*Pixel-Level*
					98.73/99.29/97.83	97.48/64.82/65.06/93.38	96.23/96.62/92.49	97.98/49.34/53.51/93.41
✓					99.09/99.55/98.34	97.74/66.85/66.84/93.81	97.26/97.53/93.83	97.89/50.33/54.45/93.34
	✓				98.98/99.42/98.15	97.61/65.66/65.55/93.59	96.19/96.53/92.55	98.05/49.84/53.69/93.35
		✓			99.21/99.54/98.51	97.78/65.24/65.46/93.80	96.88/97.26/93.22	98.20/50.11/54.08/93.66
			✓		99.19/99.56/98.34	97.83/66.00/66.24/93.66	97.03/97.29/93.40	98.28/50.41/54.25/93.77
				✓	98.97/99.40/98.36	97.74/65.99/65.95/93.73	97.42/97.52/94.13	98.54/51.03/54.73/94.87
✓	✓				99.32/99.67/98.72	97.95/67.40/67.44/94.14	97.52/97.64/94.32	98.01/50.71/54.61/93.76
✓	✓	✓			99.64/99.78/99.23	98.25/67.69/67.97/94.51	98.15/98.35/95.03	98.43/50.98/54.92/94.02
✓	✓	✓	✓		99.64/99.79/**99.33**	**98.36**/**69.00**/**68.85**/94.57	98.21/98.34/95.13	98.61/51.01/54.97/94.31
✓	✓	✓	✓	✓	**99.74**/**99.82**/99.30	98.34/68.97/68.64/**94.81**	**98.89**/**99.02**/**96.47**	**98.92**/**53.22**/**56.29**/**95.28**

TABLE 12
Integrating Dinomaly components on ViTAD (%). Models are trained for 20,000 iterations.

Method&Data	Components	I-ROC	P-ROC	P-AP	P-PRO
ViTAD [49] (MVTec-AD)	*reproduced*	98.27	97.49	58.57	91.45
	↑+NB	98.66	97.93	61.73	92.94
	↑+UA	98.56	98.01	62.20	93.14
	↑+LC	98.92	98.25	62.16	93.05
	↑+LL	98.95	98.32	**62.68**	93.21
	↑+CR	**99.25**	**98.38**	62.18	**93.48**
ViTAD [49] (VisA)	*reproduced*	90.69	97.52	39.65	83.24
	↑+all	**94.34**	**98.58**	**43.69**	**88.69**

noise with *scale* to control the noise magnitude). While both Dropout and feature jitter serve as effective noise injectors, Dropout proves to be more robust to hyperparameter and more elegant as it introduces no additional modules. Furthermore, applying dropout in the decoder (Noisy $\mathcal{D}$: 99.02%) or both locations (Noisy $\mathcal{B}+\mathcal{D}$: 99.26%) harms the performance, validating our theoretical interpretation: the Noisy Bottleneck functions as a information constraint rather than a generic regularization technique.

As shown in Table 14, Unfocused Linear Attention consistently outperforms alternatives including standard convolutions and Softmax Attention. Linear Attention surpasses vanilla Softmax Attention using less computation, validating our counterintuitive approach of leveraging attention's inability to focus as a desirable property.

We quantitatively examine different reconstruction schemes presented in Fig. 5. Table 15 shows that group-to-group Loose Constraint outperforms layer-to-layer supervision. our 2-group Loose Constraint strategy (grouping into shallow and deep semantic levels) achieves optimal performance (99.64% I-AUROC), outperforming both single-group and more granular 4-group strategies. This supports our hypothesis that appropriate semantic grouping provides the decoder with sufficient degrees of freedom without losing important hierarchical information.

Furthermore, we visualize encoder patch features using t-SNE [79] on MVTec-AD. As shown in Fig. 8, without recentering, anomalous patches from one class often overlap

TABLE 13
Ablation of Noisy Bottleneck on MVTec-AD (%). Models are trained for 10,000 iterations with ViT-B, UA, and LC.

Noise type	Noise rate	I-ROC	P-ROC	P-AP	P-PRO
None	0	99.24	97.87	65.76	93.95
Synthetic Anomaly	DRAEM	99.42	97.69	62.40	93.65
Noisy $\mathcal{B}$ (Feature Jitter)	s=5	98.94	97.64	65.23	92.88
	s=10	99.27	98.05	67.85	94.03
	s=20	99.58	98.13	68.16	94.22
	s=40	99.56	98.05	67.81	94.53
Noisy $\mathcal{B}$ (Dropout)	p=0.1	99.56	98.25	67.79	94.28
	p=0.2	99.64	98.25	67.69	94.51
	p=0.3	99.65	**98.29**	**68.28**	94.49
	p=0.4	**99.72**	**98.29**	67.76	94.47
	p=0.5	99.63	98.25	67.54	**95.01**
Noisy $\mathcal{D}$	p=0.2	99.02	97.71	66.00	93.75
Noisy $\mathcal{B} + \mathcal{D}$	p=0.2	99.26	98.06	67.82	94.25

TABLE 14
Ablation of Unfocused Attention on MVTec-AD (%). Models are trained for 10,000 iterations with ViT-B, NB, and LC.

Spatial Mixer	I-ROC	P-ROC	P-AP	P-PRO
ConvBlock 1x1	99.45	98.05	65.35	94.01
ConvBlock 3x3	99.43	98.02	65.85	94.14
ConvBlock 5x5	99.42	98.07	67.25	94.29
Softmax Attention	99.49	98.17	67.65	94.33
Unfocused Linear Attention	**99.64**	**98.25**	**67.97**	**94.51**

with normal patches from another class in the feature space. Such cases cause cross-class confusion: the decoder learns to reconstruct anomalous patterns of one class via the normal samples of the other classes. After encoder features become well-separated across different classes. Meanwhile, samples within the same class are recentered by their shared class token, preserving the discriminability between normal and anomalous patterns within each class.

4.4 Scaling Properties of Dinomaly2

We further investigated the scaling behaviors of UAD models. *(1) Model Size Scaling.* Unlike previous UAD methods that often reported diminishing or negative returns from larger models [15], [49], Dinomaly2 demonstrates clear positive scaling behavior across model sizes. As shown in Table 16, scaling from ViT-S (37.3M parameters, including bottleneck and decoder) to ViT-B (146.6M) and further to ViT-L (410.7M) consistently yields steady performance across all datasets and metrics.

Importantly, this scaling is achieved with reasonable and flexible computational cost. Dinomaly2-S already produced state-of-the-art results while maintaining a high inference speed of 253 images/second on a consumer-level GPU (NVIDIA RTX4090). While Dinomaly2-L requires higher computational resources, its throughput of 45 images/second remains practical for many real-world applications. This scalability enables practitioners to choose appropriate model sizes based on their performance-efficiency trade-offs.

(2) Input Resolution Scaling. Table 17 reveals a sharp contrast between Dinomaly2 and existing methods regarding input size scaling. Previous methods such as RD4AD and ReContrast suffer performance degradation when input resolution increases (e.g., RD4AD drops from 94.6% to 91.9%

TABLE 15
Ablation of Loose Constraint on MVTec-AD (%). Models are trained for 10,000 iterations with ViT-B, NB, and UA.

Constraints	I-ROC	P-ROC	P-AP	P-PRO
layer2layer (last 1)	98.80	97.15	60.16	92.67
layer2layer (dense, every 1)	99.32	97.95	67.40	94.14
layer2layer (sparse, every 2)	99.47	98.01	67.15	94.32
layer2layer (sparse, every 4)	99.43	97.94	66.03	94.10
group2group (1 group)	99.61	98.19	65.22	94.15
group2group (2 group)	**99.64**	**98.25**	**67.69**	**94.51**
group2group (4 group)	99.47	98.11	67.64	94.32

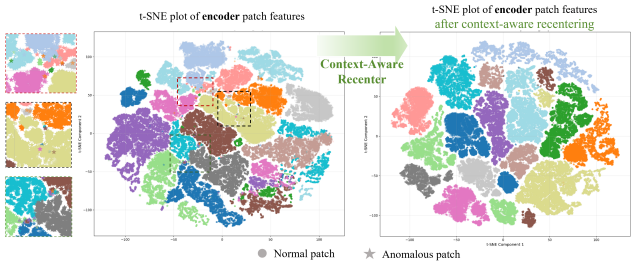

Fig. 8. t-SNE visualization of encoder patch features on MVTec-AD. Different colors represent different classes.

I-AUROC when scaling from 256^2 to 384^2). In contrast, Dinomaly2 benefits from larger input sizes, improving from 99.6% at 256^2 to 99.8% at 448^2 for image-level detection. More importantly, pixel-level performance shows substantial gains (from 65.2% AP at 280^2 to 70.3% AP at 448^2), demonstrating that our method effectively leverages higher resolution for more precise anomaly localization. Furthermore, Fig. 9 demonstrates that our approach achieves superior performance-computation trade-offs by compound scaling of model size and input size.

(3) Pre-trained Foundation Models. The representation quality of frozen backbones plays a pivotal role in unsupervised anomaly detection performance. To systematically investigate this relationship, we conduct extensive experiments across diverse pre-training paradigms. DeiT [82] employs supervised learning on ImageNet through knowledge distillation from CNNs. The masked image modeling (MIM) family includes MAE [80] (pixel-level reconstruction), BEiTv2 [83] (VQ-GAN and CLIP token prediction), and D-iGPT [84] (CLIP feature prediction). Contrastive learning (CL) approaches are represented by MOCOv3 [85] and DINO [81], which optimize feature similarity for augmented views of the same image. The most advanced models combine both strategies: iBot [50], DINOv2 [51], and its register-enhanced variant DINOv2-R [67]. Additionally, we evaluate two recently released foundation models: TIPS [86], which incorporates spatial-aware text-image pre-training, and DINOv3 [52], the latest successor to DINOv2.

An important consideration emerges regarding input resolution generalization. Most models are pre-trained with the image resolution of 224^2, except that DINOv2, DINOv2-R, TIPS, and DINOv3 have extra a high-resolution training phase with 518^2. Directly using the pre-trained weights on a different resolution for UAD without fine-tuning the parameters of enocder like other supervised tasks can cause generalization problems. Therefore, by default, we still keep the feature size of all compared models to 28×28, *i.e.*,

TABLE 16
Model size scaling analysis across model sizes on MVTecAD, VisA, and Real-IAD (%). FPS (Troughoutput, image per second) is measured on NVIDIA GeForce RTX 3090/4090 with batch size=16.

	Arch.	Params	MACs	FPS 3090/4090	MVTec-AD			VisA			Real-IAD			
					I-ROC	P-ROC	P-PRO	I-ROC	P-ROC	P-PRO	O-ROC	I-ROC	P-ROC	P-PRO
Dinomaly2-S	ViT-S	37.3M	26.2G	153.6/253.0	99.62	98.35	94.91	98.69	98.96	94.98	93.71	90.25	98.99	95.32
Dinomaly2-B	ViT-B	146.6M	103.5G	58.1/103.8	99.80	98.35	94.76	99.15	99.06	95.54	94.30	91.57	99.16	95.99
Dinomaly2-L	ViT-L	410.7M	272.9G	24.2/45.2	99.89	98.55	95.08	99.29	99.15	95.38	94.93	92.07	99.24	96.24

TABLE 17
Input resolution scaling comparison on MVTec-AD (%). †: default of the method. Note that previous methods often yield degradation with larger input sizes.

Method	Input Size	I-ROC	P-ROC	**P-AP**	**P-PRO**
RD4AD [8]	256×256†	**94.6**	**96.1**	48.6	**91.1**
	320×320	93.2	95.7	**55.1**	**91.1**
	384×384	91.9	94.9	52.1	90.8
ReContrast [11]	256×256†	**98.3**	**97.1**	60.2	93.2
	320×320	98.2	96.8	**61.8**	**93.3**
	384×384	95.2	96.5	57.7	92.6
ViTAD [49]	256×256†	**98.3**	**97.7**	55.3	91.4
	320×320	**98.3**	97.6	61.3	**92.4**
	384×384	97.8	97.5	**62.5**	**92.4**
Dinomaly2-B	280×280	**99.8**	98.3	66.7	94.2
	336×336	**99.8**	98.3	68.0	94.4
	392×392†	**99.8**	**98.4**	69.3	94.8
	448×448	**99.8**	**98.4**	**70.3**	**95.1**

TABLE 18
Comprehensive evaluation of pre-trained Vision Transformer foundations on MVTec-AD (%). The patch size of DINOv2, DINOv2-R, and TIPS is 14x14; others are 16x16. MIM: masked image modeling. CL: contrastive learning. VLM: vision language modeling. Dv1 and Dv2 denote Dinomaly and Dinomaly2.

Foundation Model	Pre-train Type	ImageNet Linear-prob	Method	Image Size	I-ROC	P-ROC	P-AP	P-PRO
MAE [80]	MIM	68.0	ViTAD	256^2	95.3	97.4	53.0	90.6
DINO [81]	CL	78.2	ViTAD	256^2	98.3	97.7	55.3	91.4
DINOv2 [51]	CL+MIM	84.5	ViTAD	224^2	98.7	97.6	55.3	92.0
DINOv2-R [67]	CL+MIM	84.8	ViTAD	224^2	98.5	97.4	54.5	92.1
DeiT [82]	Supervised	-	Dv1	224^2	97.65	97.80	62.58	89.98
MAE [80]	MIM	68.0	Dv1	224^2	97.25	97.78	63.00	90.95
BEiTv2 [83]	MIM	80.1	Dv1	224^2	97.70	97.61	59.79	90.10
D-iGPT [84]	MIM	80.5	Dv1	224^2	99.21	98.08	60.05	91.78
MOCOv3 [85]	CL	76.7	Dv1	224^2	98.74	98.05	63.36	91.13
DINO [81]	CL	78.2	Dv1	224^2	99.20	98.16	64.16	92.02
iBOT [50]	CL+MIM	79.5	Dv1	224^2	99.31	98.25	64.01	91.68
DINOv2 [51]	CL+MIM	84.5	Dv1	224^2	99.26	97.95	62.27	92.80
DINOv2-R [67]	CL+MIM	84.8	Dv1	224^2	99.34	98.09	63.04	92.59
DeiT [82]	Supervised	-	Dv1	448^2	98.19	97.93	68.98	91.45
DeiT [82]	Supervised	-	Dv2	448^2	98.41	97.99	68.31	91.85
D-iGPT [84]	MIM	80.5	Dv1	448^2	98.75	98.30	65.77	92.34
MOCOv3 [85]	CL	76.7	Dv1	448^2	98.47	98.52	70.99	92.83
DINO [81]	CL	78.2	Dv1	448^2	98.97	98.52	70.89	93.48
TIPS [86]	VLM	81.8	Dv1	392^2	99.53	98.39	67.17	94.62
TIPS [86]	VLM	81.8	Dv2	392^2	99.69	98.18	64.63	93.81
iBOT [50]	CL+MIM	79.5	Dv1	448^2	99.22	98.60	70.78	93.33
iBOT [50]	CL+MIM	79.5	Dv2	448^2	99.38	98.27	68.76	91.73
DINOv2 [51]	CL+MIM	84.5	Dv1	392^2	99.55	98.26	68.35	94.83
DINOv2 [51]	CL+MIM	84.5	Dv2	392^2	99.83	98.46	69.83	95.16
DINOv2-R [67]	CL+MIM	84.8	Dv1	392^2	99.60	98.35	69.29	94.79
DINOv2-R [67]	CL+MIM	84.8	Dv2	392^2	99.80	98.35	69.28	94.76
$\text{DINOv3}_{\text{ViT-S}}$ [52]	CL+MIM	-	Dv2	448^2	99.50	98.00	70.78	94.18
$\text{DINOv3}_{\text{ViT-B}}$ [52]	CL+MIM	-	Dv2	448^2	99.82	98.40	71.88	94.90
$\text{DINOv3}_{\text{ViT-L}}$ [52]	CL+MIM	-	Dv2	448^2	99.85	98.86	73.04	95.58

the input size is 392^2 for ViT-B/14 and 448^2 for ViT-B/16. Additionally, we train our model with the low-resolution input size of 224^2.

The results are presented in Table 18. Within Dinomaly or Dinomaly2, nearly all foundation models yields SOTA-level results with I-AUROC higher than 98%. Employing MAE produces the worst results, which aligns with known limitations of masked auto-encoders for downstream tasks without fine-tuning. This finding corroborates previous observations in ViTAD [49], where MAE as backbone yielded

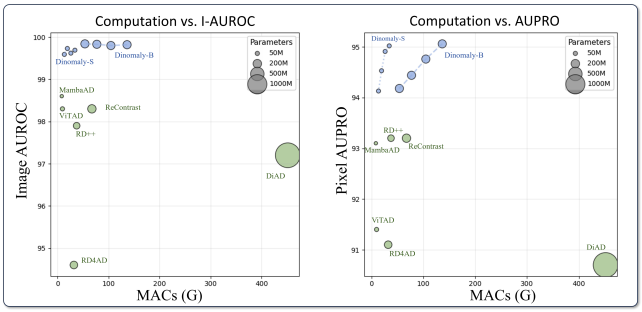

Fig. 9. Computational trade-off comparing with SOTA UAD methods on MVTec-AD. MACs: Multiply accumulate operations.

suboptimal results (95.3%), attributed to weak semantic expression from its pixel-level reconstruction objective. Models combining CL and MIM consistently outperform single-strategy counterparts, demonstrating the synergistic benefits of multi-objective pre-training. Most foundations' pixel-level performance benefit from a higher resolution; however, for the model pretrained on 224^2, image-level performance may suffer from it. Because some methods, *i.e.*, D-iGPT, DINO, and iBOT, produce similar results to DINOv2 in 224^2, suggesting that they could achieve superior performance if pre-trained at higher resolutions.

Remarkably, we observe a strong correlation ($\mathcal{R}^2$=0.91) between anomaly detection performance and ImageNet linear-probing accuracy of the foundation model. This suggests that better visual representations translate directly to improved anomaly detection capability, providing a simple heuristic for backbone selection. The recent release of DINOv3 [52] provides a compelling real-world validation of this predictive principle. As the most advanced foundation model for dense prediction, DINOv3 achieves exceptional pixel-level localization performance in our framework: 71.9/94.9% in P-AP/AUPRO with ViT-B and 73.0/95.6% with ViT-L.

5 Conclusion

We presented Dinomaly2, a minimalist yet universal framework for unsupervised anomaly detection that fundamentally reimagines how UAD systems should be designed. We present five key elements in Dinomaly2 that collectively achieve superior performance across diverse data modalities, task settings, and application domains without complex architectural designs. Extensive experiments on various benchmarks demonstrate our superiority over previous specialized methods. The combination of architectural simplicity, universal applicability, and superior performance positions Dinomaly2 as a practical framework for the full spectrum of real-world anomaly detection.

ACKNOWLEDGMENTS

The authors acknowledge supports from National Key Research and Development Program of China (2022YFC2405200, 2025YFC2426300), National Natural Science Foundation of China (U22A2051, 82027807, 82572314, 62271246, 82572313, 62506036), Tsinghua-Foshan Innovation Special Fund (2021THFS0104), the Institute for Intelligent Healthcare, Tsinghua University (2022ZLB001), the Science and Technology Commission of Shanghai Municipality (Nos.24511104100), and the Institute of Digital Medicine, City University of Hong Kong (Project 9229503).

REFERENCES

[1] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9592–9600. 1, 3, 8, 9

[2] C. Wang, W. Zhu, B.-B. Gao, Z. Gan, J. Zhang, Z. Gu, S. Qian, M. Chen, and L. Ma, "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*, 2024, pp. 22 883–22 892. 1, 3, 7, 8, 9

[3] T. Xiang, Y. Zhang, Y. Lu, A. Yuille, C. Zhang, W. Cai, and Z. Zhou, "Exploiting structural consistency of chest anatomy for unsupervised anomaly detection in radiography images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 9, pp. 6070–6081, 2024. 1

[4] J. Guo, S. Lu, L. Jia, W. Zhang, and H. Li, "Encoder-decoder contrast for unsupervised anomaly detection in medical images," *IEEE transactions on medical imaging*, vol. 43, no. 3, pp. 1102–1112, 2023. 1

[5] A. B. Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," *Expert Systems with Applications*, vol. 91, pp. 480–491, 2018. 1

[6] Y. Yao, X. Wang, M. Xu, Z. Pu, Y. Wang, E. Atkins, and D. J. Crandall, "DoTA: Unsupervised detection of traffic anomaly in driving videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 444–459, 2022. 1

[7] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 318–14 328. 1, 3, 10, 11

[8] H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9737–9746. 1, 3, 6, 9, 10, 11, 14

[9] Z. You, L. Cui, Y. Shen, K. Yang, X. Lu, Y. Zheng, and X. Le, "A unified model for multi-class anomaly detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4571–4584, 2022. 1, 3, 4, 5, 9, 10, 11

[10] Y. Zhao, "OmniAL: A unified CNN framework for unsupervised anomaly localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3924–3933. 1, 4

[11] J. Guo, S. Lu, L. Jia, W. Zhang, and H. Li, "Recontrast: Domain-specific anomaly detection via contrastive reconstruction," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023, pp. 10 721–10 740. 1, 3, 4, 7, 8, 9, 10, 11, 14

[12] R. Lu, Y. Wu, L. Tian, D. Wang, B. Chen, X. Liu, and R. Hu, "Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection," *Advances in Neural Information Processing Systems*, vol. 36, pp. 8487–8500, 2023. 1, 3, 5

[13] H. Yin, G. Jiao, Q. Wu, B. F. Karlsson, B. Huang, and C. Y. Lin, "LafitE: Latent diffusion model with feature editing for unsupervised multi-class anomaly detection," *arXiv preprint arXiv:2307.08059*, 2023. 1, 3, 5

[14] H. He, J. Zhang, H. Chen, X. Chen, Z. Li, X. Chen, Y. Wang, C. Wang, and L. Xie, "A diffusion-based framework for multi-class anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, 2024, pp. 8472–8480. 1, 3, 9, 11

[15] H. He, Y. Bai, J. Zhang, Q. He, H. Chen, Z. Gan, C. Wang, X. Li, G. Tian, and L. Xie, "MambaAD: Exploring state space models for multi-class unsupervised anomaly detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 71 162–71 187, 2024. 1, 3, 8, 9, 10, 13

[16] E. Horwitz and Y. Hoshen, "Back to the feature: classical 3D features are (almost) all you need for 3D anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2968–2977. 1, 3, 10

[17] Y. Wang, J. Peng, J. Zhang, R. Yi, Y. Wang, and C. Wang, "Multimodal industrial anomaly detection via hybrid fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 8032–8041. 1, 3, 10

[18] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "WinCLIP: Zero-/few-shot anomaly classification and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 606–19 616. 2, 3, 8, 10

[19] J. Zhu and G. Pang, "Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 17 826–17 836. 2, 4, 10

[20] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017. 2, 3

[21] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 392–408. 3, 8, 9

[22] L. Fan, D. Fan, Z. Hu, Y. Ding, D. Di, K. Yi, M. Pagnucco, and Y. Song, "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*, 2025, pp. 25 518–25 527. 3, 7, 8, 9

[23] P. Bergmann, X. Jin, D. Sattlegger, and C. Steger, "The MVTec 3D-AD dataset for unsupervised 3D anomaly detection and localization," in *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, 2022, pp. 202–213. 3, 7, 8, 9, 10

[24] W. Li, B. Zheng, X. Xu, J. Gan, F. Lu, X. Li, N. Ni, Z. Tian, X. Huang, S. Gao *et al.*, "Multi-sensor object anomaly detection: Unifying appearance, geometry, and internal properties," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 9984–9993. 3, 7, 8, 10

[25] J. Bao, H. Sun, H. Deng, Y. He, Z. Zhang, and X. Li, "BMAD: Benchmarks for medical anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4042–4053. 3, 8

[26] G. Guan, Z. Li, Y. Ma, P. Ye, J. Cao, M.-K. Wong, V. W. S. Ho, L.-Y. Chan, H. Yan, C. Tang *et al.*, "Cell lineage-resolved embryonic morphological map reveals signaling associated with cell fate and size asymmetry," *Nature Communications*, vol. 16, no. 1, p. 3700, 2025. 3, 8, 11

[27] T. Bao, J. Chen, W. Li, X. Wang, J. Fei, L. Wu, R. Zhao, and Y. Zheng, "MIAD: A maintenance inspection dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 993–1002. 3, 8, 11

[28] P. Jin, L. Mou, G.-S. Xia, and X. X. Zhu, "Anomaly detection in aerial videos with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022. 3, 8, 11

[29] J. Guo, S. Lu, W. Zhang, F. Chen, H. Li, and H. Liao, "Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 405–20 415. 3, 9, 11

[30] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Asian conference on computer vision*. Springer, 2018, pp. 622–637. 3

[31] N. Madan, N.-C. Ristea, R. T. Ionescu, K. Nasrollahi, F. S. Khan, T. B. Moeslund, and M. Shah, "Self-supervised masked convolutional transformer block for anomaly detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 1, pp. 525–542, 2023. 3

[32] J. Hyun, S. Kim, G. Jeon, S. H. Kim, K. Bae, and B. J. Kang, "Reconpatch: Contrastive patch representation learning for industrial anomaly detection," in *Proceedings of the IEEE/CVF Winter*

Conference on Applications of Computer Vision, 2024, pp. 2052–2061. 3, 10, 11

[33] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution knowledge distillation for anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 902–14 912. 3, 6

[34] H. Zhang, Z. Wang, D. Zeng, Z. Wu, and Y.-G. Jiang, "DiffusionAD: Norm-guided one-step denoising diffusion for anomaly detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3, 10

[35] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "Cutpaste: Self-supervised learning for anomaly detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9664–9674. 3

[36] V. Zavrtanik, M. Kristan, and D. Skoč aj, "DRAEM-a discriminatively trained reconstruction embedding for surface anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8330–8339. 3, 5, 11

[37] Q. Chen, H. Luo, C. Lv, and Z. Zhang, "A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization," in *European Conference on Computer Vision*. Springer, 2024, pp. 37–54. 3, 10

[38] X. Zhang, S. Li, X. Li, P. Huang, J. Shan, and T. Chen, "DeSTSeg: Segmentation guided denoising student-teacher for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3914–3923. 3, 5, 9, 10, 11

[39] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "SimpleNet: A simple network for image anomaly detection and localization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 20 402–20 411. 3, 9, 10

[40] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A patch distribution modeling framework for anomaly detection and localization," in *International Conference on Pattern Recognition*. Springer, 2021, pp. 475–489. 3

[41] S. Lee, S. Lee, and B. C. Song, "CFA: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization," *IEEE Access*, vol. 10, pp. 78 446–78 454, 2022. 3

[42] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758. 3

[43] C. Wang, H. Zhu, J. Peng, Y. Wang, R. Yi, Y. Wu, L. Ma, and J. Zhang, "M3DM-NR: Rgb-3D noisy-resistant industrial anomaly detection via multimodal denoising," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[44] W. Zhu, L. Wang, Z. Zhou, C. Wang, Y. Pan, R. Zhang, Z. Chen, L. Cheng, B.-B. Gao, J. Zhang *et al.*, "Real-IAD D3: A real-world 2D/Pseudo-3D/3D dataset for industrial anomaly detection," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 15 214–15 223. 3

[45] Y. Li, A. Goodge, F. Liu, and C.-S. Foo, "PromptAD: Zero-shot anomaly detection using text prompts," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1093–1102. 4, 10

[46] B.-B. Gao, "MetaUAS: Universal anomaly segmentation with one-prompt meta-learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 39 812–39 836, 2024. 4, 8

[47] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. 4

[48] T. Reiss, N. Cohen, E. Horwitz, R. Abutbul, and Y. Hoshen, "Anomaly detection requires better representations," in *European Conference on Computer Vision*. Springer, 2022, pp. 56–68. 4

[49] J. Zhang, X. Chen, Y. Wang, C. Wang, Y. Liu, X. Li, M.-H. Yang, and D. Tao, "Exploring plain vit features for multi-class unsupervised visual anomaly detection," *Computer Vision and Image Understanding*, vol. 253, p. 104308, 2025. 4, 6, 8, 9, 10, 11, 12, 13, 14

[50] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "Image BERT pre-training with online tokenizer," in *International Conference on Learning Representations*, 2022. 4, 13, 14

[51] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research Journal*, pp. 1–31, 2024. 4, 13, 14

[52] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa *et al.*, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025. 4, 13, 14

[53] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014. 5

[54] A. Achille and S. Soatto, "Information dropout: Learning optimal representations through noisy computation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2897–2905, 2018. 5

[55] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 ieee information theory workshop (itw)*. Ieee, 2015, pp. 1–5. 5

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008. 5

[57] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast autoregressive transformers with linear attention," in *International conference on machine learning*. PMLR, 2020, pp. 5156–5165. 5

[58] D. Han, X. Pan, Y. Han, S. Song, and G. Huang, "Flatten transformer: Vision transformer using focused linear attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 5961–5971. 6

[59] W. Sun, Z. Qin, H. Deng, J. Wang, Y. Zhang, K. Zhang, N. Barnes, S. Birchfield, L. Kong, and Y. Zhong, "Vicinity vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 635–12 649, 2023. 6

[60] H. Wi, J. Choi, and N. Park, "Learning advanced self-attention for linear transformers in the singular value domain," *arXiv preprint arXiv:2505.08516*, 2025. 6

[61] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015. 6

[62] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Asymmetric student-teacher networks for industrial anomaly detection," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 2592–2602. 7, 8, 10

[63] S. Jezek, M. Jonak, R. Burget, P. Dvorak, and M. Skotak, "Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions," in *International Congress on Ultra Modern Telecommunications and Control Systems and Workshops*. IEEE, 2021, pp. 66–71. 8, 9

[64] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti, "Vt-adl: A vision transformer network for image anomaly detection and localization," in *International Symposium on Industrial Electronics*. IEEE, 2021, pp. 01–06. 8, 9

[65] X. Jiang, J. Liu, J. Wang, Q. Nie, K. Wu, Y. Liu, C. Wang, and F. Zheng, "Softpatch: Unsupervised anomaly detection with noisy data," *Advances in Neural Information Processing Systems*, vol. 35, pp. 15 433–15 445, 2022. 8

[66] J. Zhang, H. He, Z. Gan, Q. He, Y. Cai, Z. Xue, Y. Wang, C. Wang, L. Xie, and Y. Liu, "ADer: A Comprehensive Benchmark for Multi-class Visual Anomaly Detection," *arXiv preprint arXiv:2406.03262*, 2024. 8, 11

[67] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, "Vision transformers need registers," in *International Conference on Learning Representations*, 2024. 8, 13, 14

[68] M. Wortsman, T. Dettmers, L. Zettlemoyer, A. Morcos, A. Farhadi, and L. Schmidt, "Stable and low-precision training for large-scale vision-language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 10 271–10 298, 2023. 8

[69] H. He, J. Zhang, G. Tian, C. Wang, and L. Xie, "Learning multi-view anomaly detection," *arXiv preprint arXiv:2407.11935*, 2024. 9

[70] Y.-M. Chu, C. Liu, T.-I. Hsieh, H.-T. Chen, and T.-L. Liu, "Shape-guided dual-memory learning for 3d anomaly detection," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 6185–6194. 10

[71] Y. Cao, X. Xu, and W. Shen, "Complementary pseudo multimodal feature for point cloud anomaly detection," *Pattern Recognition*, vol. 156, p. 110761, 2024. 10

[72] Y. Cheng, Y. Cao, D. Wang, W. Shen, and W. Li, "Boosting global-local feature matching via anomaly synthesis for multi-class point cloud anomaly detection," *IEEE Transactions on Automation Science and Engineering*, 2025. 10

[73] W. Lv, Q. Su, and W. Xu, "One-for-all few-shot anomaly detection via instance-induced prompt learning," in *International Conference on Learning Representations*, 2025. 10
[74] Z. Zuo, J. Dong, Y. Wu, Y. Qu, and Z. Wu, "Clip3d-ad: Extending clip for 3d few-shot anomaly detection with multi-view images generation," *arXiv preprint arXiv:2406.18941*, 2024. 10
[75] W. Luo, Y. Cao, H. Yao, X. Zhang, J. Lou, Y. Cheng, W. Shen, and W. Yu, "Exploring intrinsic normal prototypes within a single image for universal anomaly detection," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 9974–9983. 10, 11
[76] Z. Zhang, M. Cai, H. Wang, G. Wu, T. Chai, and X. Zhu, "Costfilter-ad: Enhancing anomaly detection through matching cost filtering," in *International conference on machine learning*, 2025. 10, 11
[77] J. Zhou, M. Liu, Y. Ma, S. Jiang, and Y. Wang, "Multi-view attention guided feature learning for unsupervised surface defect detection," *IEEE/ASME Transactions on Mechatronics*, 2025. 10, 11
[78] M. K. Balwant, S. Mishra, and R. Misra, "MADViT: A vision transformer-based multilayer distillation framework for explainable anomaly detection in aerial imagery," *Expert Systems with Applications*, p. 129166, 2025. 11
[79] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008. 12
[80] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009. 13, 14
[81] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660. 13, 14
[82] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357. 13, 14
[83] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "BEiT v2: Masked image modeling with vector-quantized visual tokenizers," *arXiv preprint arXiv:2208.06366*, 2022. 13, 14
[84] S. Ren, Z. Wang, H. Zhu, J. Xiao, A. Yuille, and C. Xie, "Rejuvenating image-GPT as strong visual representation learners," in *International Conference on Machine Learning*. PMLR, 2024, pp. 42 449–42 461. 13, 14
[85] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9640–9649. 13, 14
[86] K. kokitsi Maninis, K. Chen, S. Ghosh, A. Karpur, K. Chen, Y. Xia, B. Cao, D. Salz, G. Han, J. Dlabal, D. Gnanapragasam, M. Seyedhosseini, H. Zhou, and A. Araujo, "TIPS: Text-image pretraining with spatial awareness," in *International Conference on Learning Representations*, 2025. 13, 14