Dinomaly: The *Less Is More* Philosophy in Multi-Class Unsupervised Anomaly Detection

Jia Guo[1] Shuai Lu[2] Weihang Zhang[2 †] Fang Chen[3] Huiqi Li[2 †] Hongen Liao[1,3 ✉]

[1]School of Biomedical Engineering, Tsinghua University, Beijing, China
[2]School of Information and Electronics, Beijing Institute of Technology, Beijing, China
[3]School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China

guojia.jeremy@gmail.com lushuaie@163.com zhangweihang@bit.edu.cn
chen-fang@sjtu.edu.cn huiqili@bit.edu.cn liao@tsinghua.edu.cn

Abstract

*Recent studies highlighted a practical setting of unsupervised anomaly detection (UAD) that builds a unified model for multi-class images. Despite various advancements addressing this challenging task, the detection performance under the multi-class setting still lags far behind state-of-the-art class-separated models. Our research aims to bridge this substantial performance gap. In this paper, we present Dinomaly, a minimalist reconstruction-based anomaly detection framework that harnesses pure Transformer architectures without relying on complex designs, additional modules, or specialized tricks. Given this powerful framework consisting of only Attentions and MLPs, we found four simple components that are essential to multi-class anomaly detection: (1) Scalable foundation Transformers that extracts universal and discriminative features, (2) Noisy Bottleneck where pre-existing Dropouts do all the noise injection tricks, (3) Linear Attention that naturally cannot focus, and (4) Loose Reconstruction that does not force layer-to-layer and point-by-point reconstruction. Extensive experiments are conducted across popular anomaly detection benchmarks including MVTec-AD, VisA, Real-IAD, etc. Our proposed Dinomaly achieves impressive image-level AUROC of **99.6**%, **98.7**%, and **89.3**% on the three datasets respectively, which is not only superior to state-of-the-art multi-class UAD methods, but also achieves the most advanced class-separated UAD records. Code is available at: https://github.com/guojiajeremy/Dinomaly*

Accepted by CVPR 2025. ✉ Corresponding author. † Advisors when the project initiated.

1. Introduction

Unsupervised anomaly detection (UAD) aims to detect abnormal patterns from normal images and further localize the anomalous regions. Because of the diversity of potential anomalies and their scarcity, this task is proposed to model the accessible training sets containing only normal samples as an unsupervised paradigm. UAD has a wide range of applications, e.g., industrial defect detection [3], medical disease screening [13], and video surveillance [37], addressing the difficulty of collecting and labeling all possible anomalies in these scenarios.

Conventional works on UAD build a separate model for each object category, as shown in Figure 1(a). However, this one-class-one-model setting entails substantial storage overhead for saving models [60], especially when the application scenario necessitates a large number of object classes. For UAD methods, a compact boundary of normal patterns is vital to distinguish anomalies. Once the intra-normal patterns become exceedingly complicated due to various classes, the corresponding distribution becomes challenging to measure, consequently harming the detection performance. Recently, UniAD [60] and successive studies have been proposed to train a unified model for multi-class anomaly detection (MUAD), as shown in Figure 1(b). Under this setting, the "identity mapping" that directly copies the input as the output regardless of normal or anomaly harms the performance of conventional methods [60]. This phenomenon is caused by the diversity of multi-class normal patterns that drive the network to generalize on unseen patterns.

Within two years, a number of methods have been proposed to address MUAD, such as neighbor-masked attention [60], synthetic anomalies [68], vector quantization [36], diffusion model [16, 59], and state space model (Mamba) [17]. However, there is still a non-negligible per-

Dinomaly：多类无监督异常检测中的 *Less Is More*哲学

郭佳[1] 卢帅[2] 张伟航[2†] 陈芳[3] 李慧琪[2†] 廖宏恩[1,3✉]

1清华大学 生物医学工程学院，北京，中国 [2]北京理工大学 信息与电子学院，北京，中国 [3]上海交通大学 生物医学工程学院，上海，中国

guojia.jeremy@gmail.com lushuaie@163.com zhangweihang@bit.edu.cn chen-fang@sjtu.edu.cn huiqili@bit.edu.cn liao@tsinghua.edu.cn

摘要

*Recent studies highlighted a practical setting of unsupervised anomaly detection (UAD) that builds a unified model for multi-class images. Despite various advancements addressing this challenging task, the detection performance under the multi-class setting still lags far behind state-of-the-art class-separated models. Our research aims to bridge this substantial performance gap. In this paper, we present Dinomaly, a minimalist reconstruction-based anomaly detection framework that harnesses pure Transformer architectures without relying on complex designs, additional modules, or specialized tricks. Given this powerful framework consisting of only Attentions and MLPs, we found four simple components that are essential to multi-class anomaly detection: (1) Scalable foundation Transformers that extracts universal and discriminative features, (2) Noisy Bottleneck where pre-existing Dropouts do all the noise injection tricks, (3) Linear Attention that naturally cannot focus, and (4) Loose Reconstruction that does not force layer-to-layer and point-by-point reconstruction. Extensive experiments are conducted across popular anomaly detection benchmarks including MVTec-AD, VisA, Real-IAD, etc. Our proposed Dinomaly achieves impressive image-level AUROC of **99.6**%, **98.7**%, and **89.3**% on the three datasets respectively, which is not only superior to state-of-the-art multi-class UAD methods, but also achieves the most advanced class-separated UAD records. Code is available at:* https://github.com/ guojiajeremy/Dinomaly

Accepted by CVPR 2025. ✉ Corresponding author. † Advisors when the project initiated.

1. 引言

无监督异常检测（UAD）旨在从正常图像中检测异常模式，并进一步定位异常区域。由于潜在异常的多样性及其稀缺性，该任务被提出为一种无监督范式，仅对包含正常样本的可访问训练集进行建模。UAD具有广泛的应用，例如工业缺陷检测[3]、疾病医学筛查[13]和视频监控[37]，解决了在这些场景中收集和标注所有可能异常的困难。

传统的无监督异常检测（UAD）研究通常为每个物体类别单独建立一个模型，如图1(a)所示。然而，这种“一类一模型”的设置需要大量存储空间来保存模型[60]，尤其是在应用场景涉及大量物体类别时。对于UAD方法而言，清晰的正常模式边界对于区分异常至关重要。一旦因多类别导致内部正常模式变得极其复杂，其对应的分布将难以准确度量，从而损害检测性能。近期，UniAD[60]及后续研究提出了训练统一模型进行多类别异常检测（MUAD）的方案，如图1(b)所示。在此设定下，传统方法中无论输入正常或异常都直接复制为输出的“恒等映射”会严重损害性能[60]。这种现象源于多类别正常模式的多样性——这种多样性会驱使网络对未见模式进行泛化。

在两年内，已提出多种方法来解决MUAD问题，例如邻域掩码注意力[60]、合成异常[68]、向量量化[36]、扩散模型[16, 59]以及状态空间模型（Mamba）[17]。然而，仍存在不可忽视的性

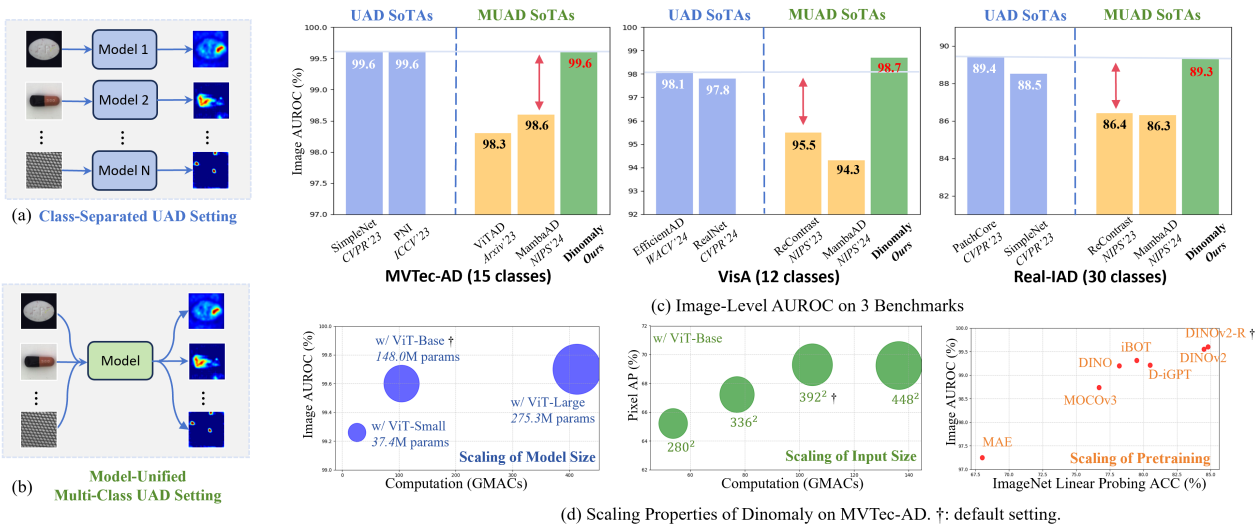


Figure 1. Setting, benchmarking, and scaling of Dinomaly. (a) Task setting of class-separated UAD. (b) Task setting of MUAD. (c) Comparison with previous SoTA methods on MVTec-AD [3], VisA [70], and Real-IAD [54]. (d) Scaling properties of Dinomaly.

formance gap between the state-of-the-art (SoTA) MUAD methods and class-separated UAD methods, restricting the practicability of implementing unified models, as shown in Figure 1(c). In addition, previous methods employ modules and architectures delicately designed, which may not be straightforward, and consequently suffer from limited universality and ease-of-use [18, 36].

In this work, we aim to catch up with the performance of class-separated anomaly detection models using a multi-class unified model. We introduce Dinomaly, a minimalist reconstruction-based UAD framework built exclusively by pure Transformer blocks [51], specifically Self-Attentions and Multi-Layer Perceptrons (MLPs). *To begin with*, we empirically investigate the scaling law of self-supervised pre-trained Vision Transformers (ViT) [12] when serving as the feature encoders for extracting reconstruction objectives. Subsequently, we introduce three straightforward yet crucial elements to address the critical identity mapping phenomenon in MUAD contexts, without increasing complexity or computational burden. *First*, as an alternative to meticulously designed pseudo anomaly and feature noise, we propose to activate the built-in Dropout in an MLP to prevent the network from restoring both normal and anomalous patterns. *Second*, we propose to leverage the "side effect" of Linear Attention (a computation-efficient counterpart of Softmax Attention) that impedes focus on local regions, thus preventing the forwarding of identical information. *Third*, previous methods adopt layer-to-layer and region-by-region reconstruction schemes, distilling a decoder that can well mimic the encoder's behavior even for anomalous regions. Therefore, we propose to loosen the reconstruction constraints by grouping multiple layers as a whole and discarding well-reconstructed regions during optimization.

To validate the effectiveness of the proposed Dinomaly under MUAD setting, we conduct extensive experiments on a number of widely used benchmarks, including MVTec AD [3] (15 classes), VisA [70] (12 classes), and Real-IAD (30 classes). As shown in Figure 1, our base-size Dinomaly achieves unprecedented image-level AUROC of **99.6**%, **98.7**%, and **89.3**% on MVTec AD, VisA, and Real-IAD, surpassing previous SoTA methods by a large margin. In addition, scalability is a key feature of Dinomaly. Further scaling up the model size maximizes performance to the fullest level of **99.8**%, **98.9**%, and **90.1**%, respectively; while scaling down parameters and input size can offer efficient solutions in computation-constrained scenarios.

2. Related Work

Multi-Class UAD. UniAD [60] first introduced multi-class anomaly detection, aiming to detect anomalies for different classes using a unified model. In this setting, conventional UAD methods often face the challenge of "identical shortcuts", where both anomaly-free and anomaly samples can be effectively recovered during inference [60]. It is believed that this phenomenon is caused by the diversity of multi-class normal patterns that drive the network to generalize on unseen patterns. This contradicts the fundamental assumption of epistemic methods. Many current researches focus on addressing this challenge [14, 31, 36, 59, 60]. UniAD [60] employs a neighbor-masked attention module and a feature-jitter strategy to mitigate these shortcuts. HVQ-Trans [36] proposes a vector quantization (VQ) Transformer model that induces large feature discrepancies

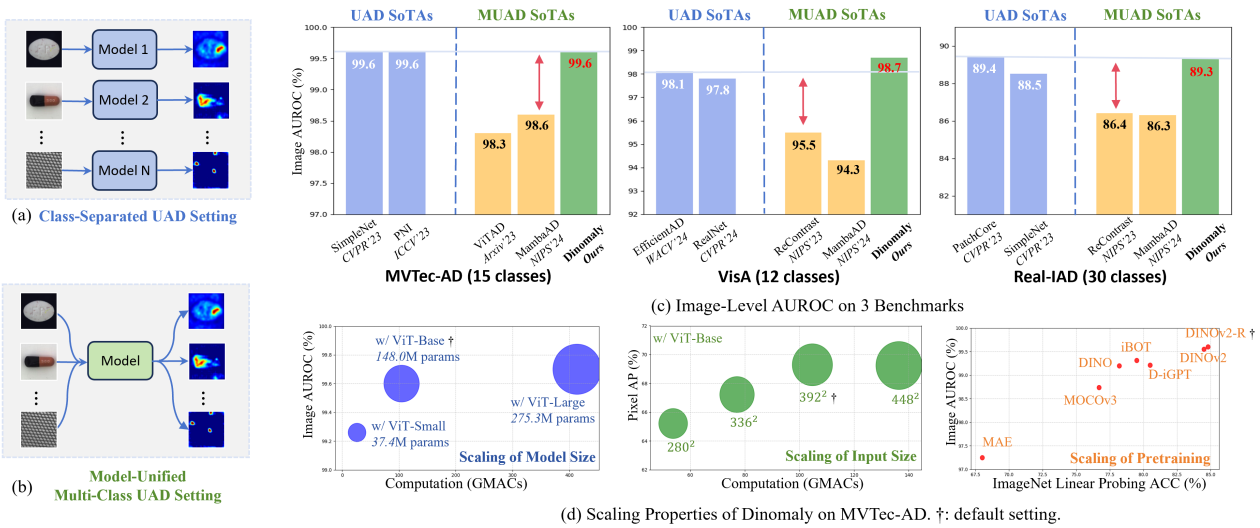


图1. Dinomaly的设置、基准测试与扩展。(a) 类别分离式无监督异常检测的任务设置。(b) 多类别无监督异常检测的任务设置。(c) 与先前在MVTec-AD [3]、VisA [70] 和 Real-IAD [54] 上的最先进方法对比。(d) Dinomaly 的缩放特性。

最先进的（SoTA）MUAD方法与类别分离的UAD方法之间存在性能差距，限制了统一模型的实际应用性，如图1(c)所示。此外，先前方法采用了精心设计的模块和架构，这可能不够直观，因此普遍性和易用性受到限制[18, 36]。

在本工作中，我们的目标是通过一个多类别统一模型来追赶类别分离异常检测模型的性能。我们提出了Dinomaly，这是一个极简主义的基于重建的统一异常检测框架，完全由纯Transformer模块构成[51]，具体包括自注意力机制和多层感知机。*To begin with*，我们实证研究了自监督预训练视觉Transformer作为特征编码器提取重建目标时的缩放规律。随后，我们引入了三个简单而关键的要素，以解决多类别统一异常检测中的关键身份映射现象，同时不增加复杂度或计算负担。*First*，作为精心设计的伪异常和特征噪声的替代方案，我们提出激活多层感知机中的内置Dropout，以防止网络同时恢复正常和异常模式。*Second*，我们提出利用线性注意力的“副作用”（这是Softmax注意力的一种计算高效替代方案）来阻碍对局部区域的聚焦，从而防止相同信息的传递。*Third*，先前的方法采用逐层和逐区域的重建方案，训练出的解码器即使对异常区域也能很好地模仿编码器的行为。因此，我们建议通过将多个层分组来放宽重建约束。

整体并在优化过程中舍弃重建良好的区域。

为了验证所提出的Dinomaly在MUAD设置下的有效性，我们在多个广泛使用的基准测试上进行了大量实验，包括MVTec AD [3]（15个类别）、VisA [70]（12个类别）和Real-IAD（30个类别）。如图1所示，我们的基础尺寸Dinomaly在MVTec AD、VisA和Real-IAD上分别实现了前所未有的图像级AUROC，达到99.6%、98.7%和89.3%，大幅超越了之前的SoTA方法。此外，可扩展性是Dinomaly的一个关键特性。进一步扩大模型尺寸可将性能最大化至99.8%、98.9%和90.1%；而缩减参数和输入尺寸则能在计算受限场景下提供高效的解决方案。

2. 相关工作

多类别无监督异常检测。UniAD [60] 首次引入了多类别异常检测，旨在通过统一模型检测不同类别的异常。在此设定下，传统无监督异常检测方法常面临"相同捷径"的挑战——即推理过程中正常样本与异常样本均能被有效重构[60]。研究认为，该现象源于多类别正常模式的多样性驱使网络对未见模式产生泛化，这与认知方法的基本假设相悖。当前许多研究聚焦于应对这一挑战[14, 31, 36, 59, 60]。UniAD [60] 采用邻域掩蔽注意力模块与特征抖动策略来缓解此类捷径问题。HVQ-Trans [36] 提出向量量化（VQ）Transformer模型，通过{v*}机制诱导显著特征差异。

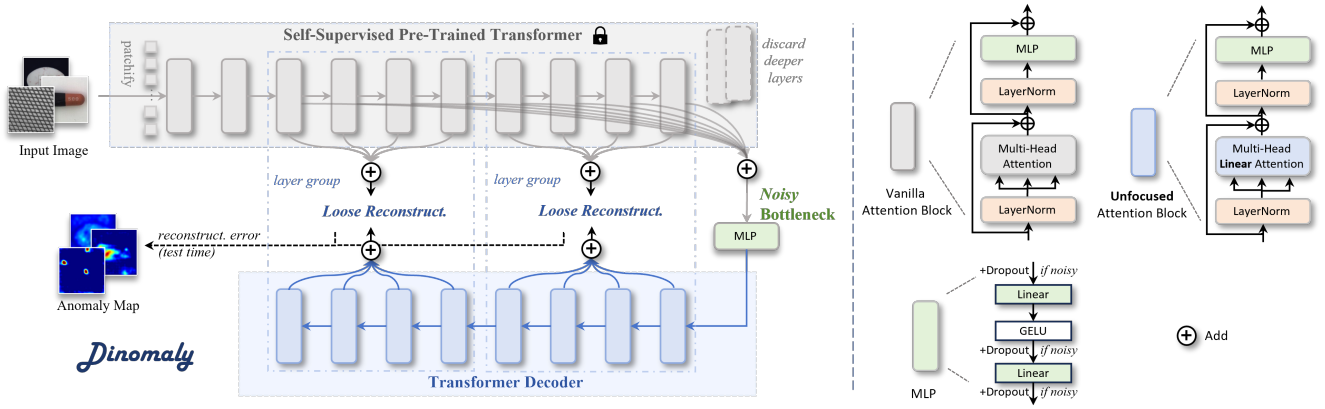


Figure 2. The framework of Dinomaly, built by simple and pure Transformer building blocks.

for anomalies. LafitE [59] utilizes a latent diffusion model and introduces a feature editing strategy to alleviate this issue. DiAD [16] also employs diffusion models to address multi-class UAD settings. OmniAL [68] focuses on anomaly localization in the unified setting, preventing identical reconstruction by using synthesized pseudo anomalies. ReContrast [14] attempted to alleviate the identity mapping by cross-reconstruction between two encoders. ViTAD [5] abstracts a unified feature-reconstruction UAD framework and employ Transformer building blocks. MambaAD [17] explores the recently proposed State Space Model (SSM), Mamba, in the context of multi-class UAD. More related works of UAD are presented in Appendix A.

3. Method

3.1. Dinomaly Framework

"What I cannot create, I do not understand"
——Richer Feynman

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. Dinomaly consists of an encoder, a bottleneck, and a reconstruction decoder, as shown in Figure 2. Without loss of generality, a pretrained ViT network [12] with 12 Transformer layers is used as the encoder, extracting informative feature maps with different semantic scales. The bottleneck is a simple MLP (a.k.a. feed-forward network, FFN) that collects the feature representations of the encoder's 8 middle-level layers. The decoder is similar to the encoder, consisting of 8 Transformer layers. During training, the decoder learns to reconstruct the middle-level features of the encoder by maximizing the cosine similarity between feature maps. During inference, the decoder is expected to reconstruct normal regions of feature maps but fails for anomalous regions as it has never seen such samples.

Foundation Transformers. Foundation models, especially ViTs [12, 33] pre-trained on large-scale datasets, serve as a basis and starting point for specific computer vision tasks. Such networks employ self-supervised learning schemes such as contrastive learning (MoCov3 [6], DINO [4]), masked image modeling (MAE [19], SimMIM [57], BEiT [40]), and their combination (iBOT [69], DINOv2 [39]), producing universal features suitable for image-level visual tasks and pixel-level visual tasks.

Because of the lack of supervision in UAD, most advanced methods adopt pre-trained networks to extract discriminative features. Recent works [28, 43, 65] have preliminarily discovered the advantage of robust and universal features of self-supervised models over domain-specific ImageNet features in anomaly detection tasks. In this paper, we pioneer the investigation of scaling behaviors in UAD models through a systematic analysis of foundational ViTs, as briefed in Figure 1(d). Our comprehensive evaluation encompasses pre-training strategies (Figure 5), model sizes (Table 4), and input resolutions (Table 5), which are detailed in section 4.4. Considering the balance of detection performance and computational efficiency, we adopt ViT-Base/14 pretrained by DINOv2-Register [7] as the encoder of Dinomaly by default.

3.2. Noisy Bottleneck.

"Dropout is all you need."

Prior studies [14, 60, 68] attribute the performance degradation of UAD methods trained on diverse multi-class samples to the "identity mapping" phenomenon; in this work, we reframe this issue as an "over-generalization" problem. Generalization ability is a merit of neural networks, allowing them to perform equally well on unseen test sets. However, generalization is not so wanted in the context of unsupervised anomaly detection that leverages the epistemic nature of neural networks. With the increas-

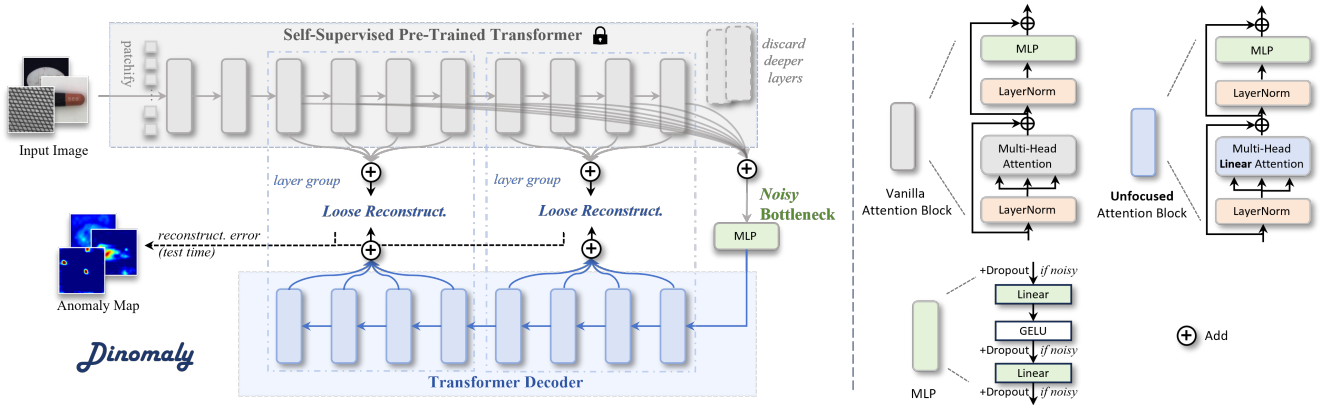


图2. Dinomaly的框架，由简单纯粹的Transformer构建模块构成。

对于异常检测。LafitE [59] 利用潜在扩散模型并引入特征编辑策略来缓解这一问题。DiAD [16] 同样采用扩散模型来处理多类别统一异常检测（UAD）设置。OmniAL [68] 专注于统一设置下的异常定位，通过使用合成的伪异常来防止完全相同的重建。ReContrast [14] 尝试通过两个编码器之间的交叉重建来缓解恒等映射问题。ViTAD [5] 抽象出一个统一的特征重建UAD框架，并采用Transformer构建模块。MambaAD [17] 在多元UAD的背景下探索了最近提出的状态空间模型（SSM）——Mamba。更多UAD相关的工作见附录A。

3. 方法

3.1. 动态异常框架

“What I cannot create, I do not understand”
——更丰富的费曼

从已知中识别异常的能力是人类与生俱来的禀赋，也是我们探索世界的重要途径。类似地，我们构建了一个基于重建的框架，该框架依赖于人工神经网络的认知特性。Dinomaly 由编码器、瓶颈层和重建解码器组成，如图2所示。在不失一般性的前提下，我们采用预训练的ViT网络[12]作为编码器，该网络包含12个Transformer层，能够提取具有不同语义尺度的信息特征图。瓶颈层是一个简单的MLP（即前馈网络，FFN），用于收集编码器中间8个层的特征表示。解码器结构与编码器相似，由8个Transformer层构成。在训练过程中，解码器通过最大化特征图间的余弦相似度来学习重建编码器的中层特征。在推理阶段，解码器能够正常重建特征图的常规区域，但对于异常区域则难以实现重建，因为{v*}从未见过这样的样本。

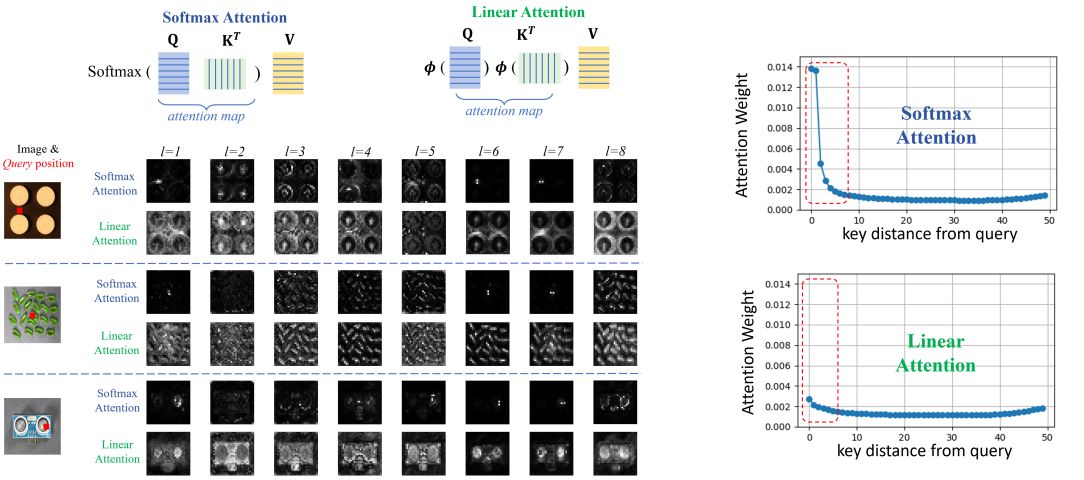
基础Transformer。基础模型，尤其是在大规模数据集上预训练的ViTs [12, 33]，为特定计算机视觉任务提供了基础和起点。这类网络采用自监督学习方案，如对比学习（MoCov3 [6]、DINO [4]）、掩码图像建模（MAE [19]、SimMIM [57]、BEiT [40]）及其组合（iBOT [69]、DINOv2 [39]），生成适用于图像级视觉任务和像素级视觉任务的通用特征。

由于无监督异常检测（UAD）中缺乏监督信号，大多数先进方法采用预训练网络来提取判别性特征。近期研究[28, 43, 65]初步发现，在异常检测任务中，自监督模型具有的鲁棒通用特征相较于领域特定的ImageNet特征更具优势。本文通过系统分析基础视觉Transformer（ViT），率先探索了UAD模型中的缩放规律，如图1(d)所示。我们的综合评估涵盖预训练策略（图5）、模型规模（表4）和输入分辨率（表5），具体细节见第4.4节。综合考虑检测性能与计算效率的平衡，我们默认采用DINOv2-Register [7]预训练的ViT-Base/14作为Dino-maly的编码器。

3.2. 噪声瓶颈。

“Dropout is all you need.”

先前的研究[14, 60, 68]将基于多样多类样本训练的UAD方法性能下降归因于“恒等映射”现象；在本工作中，我们将此问题重新定义为“过度泛化”问题。泛化能力是神经网络的优点，使其在未见过的测试集上表现同样出色。然而，在利用神经网络认知特性的无监督异常检测场景中，泛化能力并不总是有益的。随着



(a) Attention map of decoder layer l (mean of 12 heads)

(b) Attention distribution across query-key distances

Figure 3. Softmax Attention *vs.* Linear Attention. (a) Visualization of attention maps. (b) Attention distribution.

ing diversity of images and their patterns due to multi-class UAD settings, the decoder can generalize its reconstruction ability to unseen anomalous samples, resulting in the failure of anomaly detection using reconstruction error.

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, previous works propose to add perturbations as pseudo anomalies on input images [62, 67] or on encoder features [59, 60]; meanwhile, still let the decoder restore anomaly-free images or features, formulating a denoising-like framework. However, such methods employ heuristic and hand-crafted anomaly generation strategies that may not be universal across domains, datasets, and methods. In this work, we turn to leveraging the simple and elegant Dropout techniques. Since its introduction by Hinton et al. [21] in 2014 as a remedy for overfitting, Dropout has become a cornerstone element in neural architectures, including Transformers. In Dinomaly, we employ Dropout to randomly discard neural activations in an MLP bottleneck. Instead of alleviating overfitting, the role of Dropout in Dinomaly can be explained as pseudo feature anomaly that perturb normal representations, analogous to denoising auto-encoders [52, 53]. Without introducing any specific modules, this simple component inherently forces the decoder to restore normal features regardless of whether the test image contains anomalies, in turn, mitigating identical mapping.

3.3. Unfocused Linear Attention.

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

Softmax Attention is the key mechanism of Transformers, allowing the model to attend to different parts of its input token sequence. Formally, given an input sequence $\mathbf{X} \in \mathbb{R}^{N\times d}$ 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 $\mathbf{V} \in \mathbb{R}^{N\times d}$:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^Q, \mathbf{K} = \mathbf{X}\mathbf{W}^K, \mathbf{V} = \mathbf{X}\mathbf{W}^V, \quad (1)$$

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 is given as: [1]

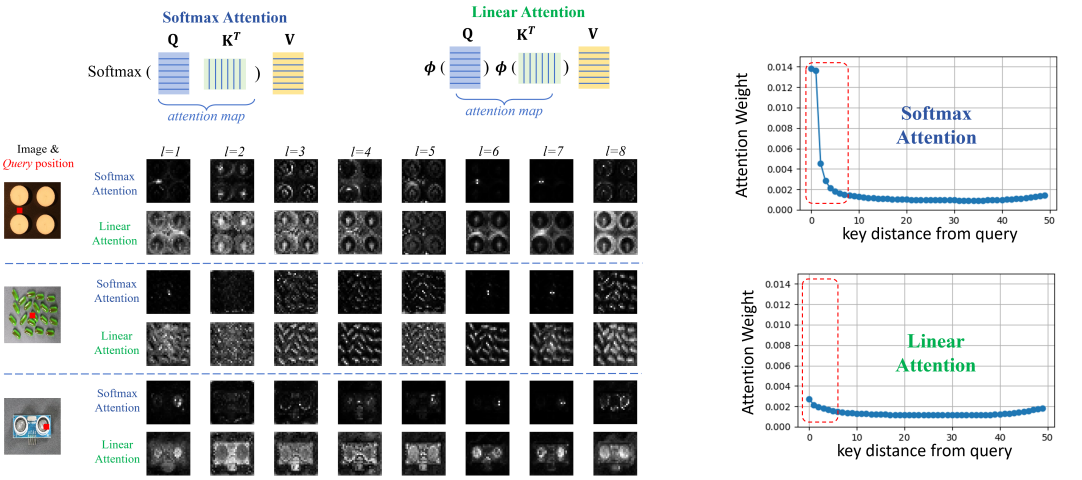
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\mathbf{Q}\mathbf{K}^T)\mathbf{V}. \quad (2)$$

Back to MUAD, previous methods [36, 60] suggest adopting Attentions instead of Convolutions because Convolutions can easily learn identical mappings. Nevertheless, both operations are in danger of forming identity mapping by over-concentrating on corresponding input locations for producing the outputs:

$$\text{Conv Kernel} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \text{Attn Map} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Is there any simple solution that prevents Attentions from attending to identical information? In Dinomaly, we turn to leverage the "unfocusing ability" of a type of Softmax-free Attention, i.e., **Linear Attention**. Linear Attention was proposed as a promising alternative to reduce the computation complexity of vanilla Softmax Attention

[1] The full form of Attention is Softmax($\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}$)$\mathbf{V}$. The constant denominator is omitted for narrative simplicity. The multi-head mechanism that concatenates multiple Attentions is also omitted.



(a) Attention map of decoder layer *l* (mean of 12 heads)

(b) Attention distribution across query-key distances

图3. Softmax注意力 *vs.* 线性注意力。(a) 注意力分布图可视化。(b) 注意力分布。

由于多类别无监督异常检测设置中图像及其模式的多样性，解码器能够将其重建能力泛化到未见过的异常样本上，导致使用重建误差进行异常检测的失败。

身份映射的直接解决方案是将“重建”转向“修复”。具体而言，先前的研究并非直接根据正常输入重建正常图像或特征，而是提出在输入图像[62, 67]或编码器特征[59, 60]上添加扰动作为伪异常；同时仍让解码器修复无异常的图像或特征，从而构建出类似去噪的框架。然而，这类方法采用启发式且手工设计的异常生成策略，可能无法跨领域、跨数据集和跨方法通用。在本工作中，我们转而利用简洁优雅的Dropout技术。自Hinton等人[21]于2014年提出Dropout作为过拟合的解决方案以来，它已成为包括Transformer在内的神经架构中的基石元素。在Dinomaly中，我们采用Dropout随机丢弃MLP瓶颈层中的神经激活。Dropout在Dinomaly中的作用并非缓解过拟合，而可解释为扰动正常表征的伪特征异常，类似于去噪自编码器[52, 53]。这一简单组件无需引入任何特定模块，便能迫使解码器无论测试图像是否包含异常都恢复出正常特征，从而缓解了身份映射问题。

3.3. 非聚焦线性注意力。

“One man’s poison is another man’s meat”

Softmax注意力是Transformer的关键机制，它使模型能够关注其不同部分输入令牌序列。形式上，给定一个长度为 N 的输入序列 $\mathbf{X} \in \mathbb{R}^{N\times d}$，注意力机制首先将其转换为三个矩阵：查询矩阵 $\mathbf{Q} \in \mathbb{R}^{N\times d}$、键矩阵 $\mathbf{K} \in \mathbb{R}^{N\times d}$ 和值矩阵 $\mathbf{V} \in \mathbb{R}^{N\times d}$：

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^Q, \mathbf{K} = \mathbf{X}\mathbf{W}^K, \mathbf{V} = \mathbf{X}\mathbf{W}^V, \quad (1)$$

其中$\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d\times d}$是可学习的参数。通过查询-键相似度计算注意力图，Softmax注意力的输出为：[1]

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\mathbf{Q}\mathbf{K}^T)\mathbf{V}. \quad (2)$$

回到MUAD，先前的方法[36, 60]建议采用注意力机制而非卷积，因为卷积容易学习恒等映射。然而，这两种操作都存在过度集中于对应输入位置以产生输出，从而形成恒等映射的风险：

$$\text{Conv Kernel} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \text{Attn Map} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

是否存在一种简单的解决方案，能防止注意力机制关注到完全相同的信息？在Dinomaly中，我们转而利用一种无需Softmax的注意力机制——即线性注意力——所具备的“失焦能力”。线性注意力被提出时，是作为降低传统Softmax注意力计算复杂度的一种有前景的替代方案。

[1]The full form of Attention is Softmax($\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}$)$\mathbf{V}$. The constant denominator is omitted for narrative simplicity. The multi-head mechanism that concatenates multiple Attentions is also omitted.

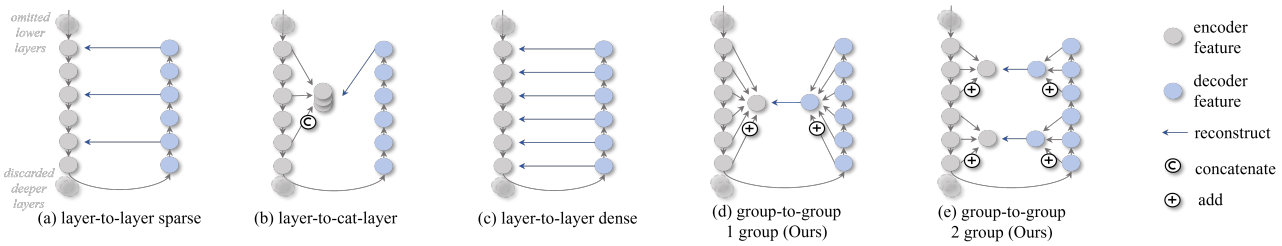


Figure 4. Schemes of reconstruction constraint. (a) Layer-to-layer (sparse). (b) Layer-to-cat-layer. (c) Layer-to-layer (dense). (d) Loose group-to-group, 1-group (Ours). (e) Loose group-to-group, 2-group (Ours).

concerning the number of tokens [26]. By substituting Softmax operation with a simple activation function $\phi(\cdot)$ (usually $\phi(x) = \text{elu}(x) + 1$), we can change the computation order from $(\mathbf{Q}\mathbf{K}^T)\mathbf{V}$ to $\mathbf{Q}(\mathbf{K}^T\mathbf{V})$. Formally, Linear Attention (LA) is given as:

$$\text{LA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = (\phi(\mathbf{Q})\phi(\mathbf{K}^T))\mathbf{V} = \phi(\mathbf{Q})(\phi(\mathbf{K}^T)\mathbf{V}) \,, \tag{3}$$

where the computation complexity is reduced to $\mathcal{O}(Nd^2)$ from $\mathcal{O}(N^2d)$. The trade-off between complexity and expressiveness is a dilemma. Previous studies [15, 48] attribute Linear Attention's performance degradation on supervised tasks to its incompetence in focusing. Due to the absence of non-linear attention reweighting by Softmax operation, Linear Attention cannot concentrate on important regions related to the query, such as foreground and neighbors. This property, however, is exactly what the reconstruction decoder favors in our contexts.

In order to probe how Attentions propagate information, we train two variants of Dinomaly using vanilla Softmax Attention or Linear Attention as the spatial mixer in the decoder and visualize their attention maps. As shown in Figure 3, Softmax Attention tends to focus on the exact region of the query, while Linear Attention spreads its attention across the whole image. This implies that Linear Attention, forced by its incompetence to focus, utilizes more long-range information to restore features at each position, reducing the chance of passing identical information of unseen patterns to the next layer during reconstruction. Of course, employing Linear Attention also benefits from less computation, free of performance drop.

3.4. Loose Reconstruction

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

Loose Constraint. Pioneers of feature-reconstruction/distillation UAD methods [10, 46] are inspired by knowledge distillation [20]. Most reconstruction-based methods distill specific encoder layers (e.g. 3 last layers of 3 ResNet stages) by the corresponding decoder layers [10, 46, 65] (Figure 4(a)) or the last decoder layer [58, 60] (Figure 4(b)). Intuitively, with more encoder-decoder feature pairs (Figure 4(c)), UAD model can utilize more information in different layers to discriminate anomalies. However, according to the intuition of knowledge distillation, the student (decoder) can better mimic the behavior of the teacher (encoder) given more layer-to-layer supervision [30], which is harmful for UAD models that detect anomalies by encoder-decoder discrepancy. This phenomenon is also embodied as identity mapping. Thanks to the top-to-bottom consistency of columnar Transformer layers, we propose to loosen the layer-to-layer constraint by adding up all feature maps of interested layers as a whole group, as shown in Figure 4(d). This scheme can be seen as loosening the layer-to-layer correspondence and providing the decoder with more degrees of freedom, so that the decoder is allowed to act much more differently from the encoder when the input pattern is unseen. Because features of shallow layers contain low-level visual characters that are helpful for precise localization, we can further group the features into the low-semantic-level group and high-semantic-level group, as shown in Figure 4(e).

Loose Loss. Following the above analysis, we also loosen the point-by-point reconstruction loss function by discarding some points in the feature map. Here, we simply borrow the hard-mining global cosine loss [14] that detaches the gradients of well-restored feature points with low cosine distance during training. Let f_E and f_D denotes (grouped) feature maps of encoder and decoder:

$$\mathcal{L}_{global-hm} = \mathcal{D}_{cos}(\mathcal{F}(f_E), \mathcal{F}(\hat{f_D})), \tag{4}$$

$$\hat{f_D}(h,w) = \begin{cases} sg(f_D(h,w))_{0.1}, & \text{if } \mathcal{D}_{cos}(f_D, f_E) < k\%_{batch} \\ f_D(h,w), & \text{else} \end{cases} \tag{5}$$

$$\mathcal{D}_{cos}(a,b) = 1 - \frac{a^T \cdot b}{\|a\| \ \|b\|}, \tag{6}$$

where $\mathcal{D}_{cos}$ denotes cosine distance, $\mathcal{F}(\cdot)$ denotes flatten operation, $f_D(h,w)$ represents the feature point at (h,w), and $sg(\cdot)_{0.1}$ denotes shrink the gradient to one-tenth of the original [2]. $\mathcal{D}_{cos}(f_D(h,w), f_E(h,w)) < k\%_{batch}$ selects $k\%$ feature points with smaller cosine distance within a batch for gradient shrinking. Total loss is the average $\mathcal{L}_{global-hm}$ of all encoder-decoder feature pairs.

[2]Complete stop-gradient causes optimization instability occasionally.

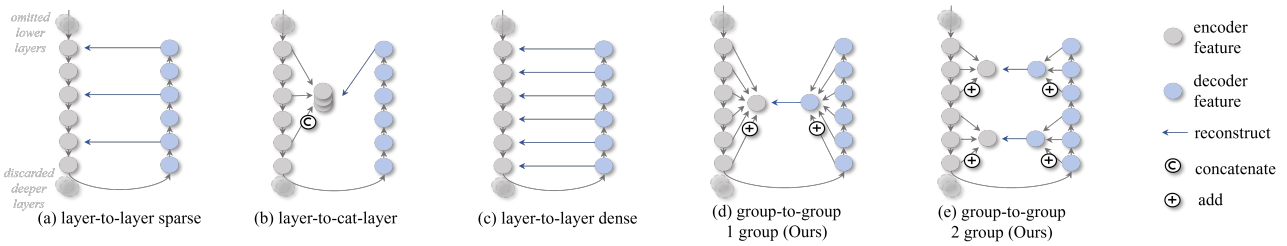


图4. 重建约束方案示意图。(a) 逐层重建（稀疏）。(b) 层到类别层。(c) 逐层重建（密集）。(d) 宽松组到组，1组（本文方法）。(e) 宽松组到组，2组（本文方法）。

关于令牌数量[26]。通过将Softmax操作替换为简单的激活函数$\phi(\cdot)$ (通常是$\phi(x) = \text{elu}(x) + 1$)，我们可以将计算顺序从$(\mathbf{QK}^T)\mathbf{V}$改为$\mathbf{Q}(\mathbf{K}^T\mathbf{V})$。形式上，线性注意力（LA）的表达式为:

$$\text{LA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = (\phi(\mathbf{Q})\phi(\mathbf{K}^T))\mathbf{V} = \phi(\mathbf{Q})(\phi(\mathbf{K}^T)\mathbf{V}), \tag{3}$$

其中计算复杂度从$\mathcal{O}(N^2d)$降低到$\mathcal{O}(Nd^2)$。复杂度与表达能力之间的权衡是一个两难问题。先前的研究[15, 48]将线性注意力在监督任务上的性能下降归因于其聚焦能力的不足。由于缺少Softmax操作带来的非线性注意力重加权机制，线性注意力无法聚焦于与查询相关的重要区域（例如前景和邻近区域）。然而这一特性恰恰是我们场景中重建解码器所需要的。

为了探究注意力机制如何传播信息，我们训练了两种Dinomaly变体，分别在解码器中使用标准Softmax注意力或线性注意力作为空间混合器，并可视化它们的注意力图。如图3所示，Softmax注意力倾向于聚焦于查询的精确区域，而线性注意力则将注意力分散到整个图像。这表明线性注意力因其聚焦能力的不足，被迫利用更多长程信息来恢复每个位置的特征，从而降低了在重建过程中将未见模式的相同信息传递到下一层的可能性。当然，采用线性注意力还能受益于更少的计算量，且不会导致性能下降。

3.4. 松散重建

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

宽松约束。特征重建/蒸馏的无监督异常检测方法先驱[10, 46]受到知识蒸馏[20]的启发。大多数基于重建的方法通过对应的解码器层[10, 46, 65]（图4(a)）或最终解码器层[58, 60]（图4(b)）来蒸馏特定编码器层（例如ResNet三个阶段中的最后3层）。直观而言，通过使用更多编码器-解码器特征对（图4(c)），无监督异常检测模型能够更充分地利用在不同层中利用更多信息来区分异常。然而，根据知识蒸馏的直觉，在更多层间监督下，学生（解码器）能更好地模仿教师（编码器）的行为[30]，这对于依赖编码器-解码器差异检测异常的UAD模型是有害的。这一现象也体现为恒等映射。得益于柱状Transformer层自上而下的一致性，我们提出通过将所有关注层的特征图作为一个整体组进行叠加来放松层间约束，如图4(d)所示。该方案可视为松动了层间对应关系，为解码器提供更多自由度，从而允许解码器在面对未见输入模式时表现出与编码器更大的差异。由于浅层特征包含有助于精确定位的低层次视觉特征，我们可进一步将特征分组为低语义级组和高语义级组，如图4(e)所示。

松散损失。根据上述分析，我们通过丢弃特征图中的部分点，进一步放宽了点对点重建损失函数。此处，我们直接借鉴了硬挖掘全局余弦损失[14]的方法，该损失在训练过程中会剥离余弦距离较小、恢复良好的特征点的梯度。令f_E和f_D分别表示编码器与解码器的（分组）特征图:

$$\mathcal{L}_{global-hm} = \mathcal{D}_{cos}(\mathcal{F}(f_E), \mathcal{F}(\hat{f_D})), \tag{4}$$

$$\hat{f_D}(h,w) = \begin{cases} sg(f_D(h,w))_{0.1}, & \text{if } \mathcal{D}_{cos}(f_D, f_E) < k\%_{batch} \\ f_D(h,w), & \text{else} \end{cases} \tag{5}$$

$$\mathcal{D}_{cos}(a,b) = 1 - \frac{a^T \cdot b}{\|a\| \ \|b\|}, \tag{6}$$

其中$\mathcal{D}_{cos}$表示余弦距离，$\mathcal{F}(\cdot)$表示展平操作，$f_D(h,w)$代表位于(h,w)处的特征点，$sg(\cdot)_{0.1}$表示将梯度缩小至原始[2]的十分之一。$\mathcal{D}_{cos}(f_D(h,w), f_E(h,w)) < k\%_{batch}$会在批次内选取余弦距离较小的$k$%特征点进行梯度收缩。总损失是所有编码器-解码器特征对$\mathcal{L}_{global-hm}$的平均值。

[2]Complete stop-gradient causes optimization instability occasionally.

Table 1. Performance under **multi-class** UAD setting (%). †: method designed for MUAD. Dinomaly↑: training schedule is scaled up to 20,000, 20,000, and 100,000 iterations (original: 10,000/10,000/50,000).

Dateset	Method	Image-level			Pixel-level			
		AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
MVTec-AD [3]	RD4AD [10]	94.6	96.5	95.2	96.1	48.6	53.8	91.1
	SimpleNet [34]	95.3	98.4	95.8	96.9	45.9	49.7	86.5
	DeSTSeg [67]	89.2	95.5	91.6	93.1	54.3	50.9	64.8
	UniAD [60]†	96.5	98.8	96.2	96.8	43.4	49.5	90.7
	ReContrast [14]†	98.3	99.4	97.6	97.1	60.2	61.5	93.2
	DiAD [18]†	97.2	99.0	96.5	96.8	52.6	55.5	90.7
	ViTAD [65]†	98.3	99.4	97.3	97.7	55.3	58.7	91.4
	MambaAD [17]†	98.6	99.6	97.8	97.7	56.3	59.2	93.1
	Dinomaly (Ours)	**99.6**	**99.8**	**99.0**	**98.4**	**69.3**	**69.2**	**94.8**
	Dinomaly↑	99.7	99.8	99.2	98.4	69.3	69.2	94.7
VisA [70]	RD4AD [10]	92.4	92.4	89.6	98.1	38.0	42.6	91.8
	SimpleNet [34]	87.2	87.0	81.8	96.8	34.7	37.8	81.4
	DeSTSeg [67]	88.9	89.0	85.2	96.1	39.6	43.4	67.4
	UniAD [60]†	88.8	90.8	85.8	98.3	33.7	39.0	85.5
	ReContrast [14]†	95.5	96.4	92.0	98.5	47.9	50.6	91.9
	DiAD [18]†	86.8	88.3	85.1	96.0	26.1	33.0	75.2
	ViTAD [65]†	90.5	91.7	86.3	98.2	36.6	41.1	85.1
	MambaAD [17]†	94.3	94.5	89.4	98.5	39.4	44.0	91.0
	Dinomaly (Ours)	**98.7**	**98.9**	**96.2**	**98.7**	**53.2**	**55.7**	**94.5**
	Dinomaly↑	98.9	99.0	96.4	98.8	53.8	55.8	94.5
Real-IAD [54]	RD4AD [10]	82.4	79.0	73.9	97.3	25.0	32.7	89.6
	SimpleNet [34]	57.2	53.4	61.5	75.7	2.8	6.5	39.0
	DeSTSeg [67]	82.3	79.2	73.2	94.6	37.9	41.7	40.6
	UniAD [60]†	83.0	80.9	74.3	97.3	21.1	29.2	86.7
	ReContrast [14]†	86.4	84.2	77.4	97.8	31.6	38.2	91.8
	DiAD [18]†	75.6	66.4	69.9	88.0	2.9	7.1	58.1
	ViTAD [65]†	82.7	80.2	73.7	97.2	24.3	32.3	84.8
	MambaAD [17]†	86.3	84.6	77.0	98.5	33.0	38.7	90.5
	Dinomaly (Ours)	**89.3**	**86.8**	**80.2**	**98.8**	**42.8**	**47.1**	**93.9**
	Dinomaly↑	89.5	86.9	80.4	98.9	43.3	47.4	94.2

4. Experiments

4.1. Experimental Settings

Datasets. **MVTec-AD** [3] contains 15 objects (5 texture classes and 10 object classes) with a total of 3,629 normal images as the training set and 1,725 images as the test set (467 normal, 1,258 anomalous). **VisA** [70] contains 12 objects. Training and test sets are split following the official splitting, resulting in 8,659 normal images in the training set and 2,162 images in the test set (962 normal, 1,200 anomalous). **Real-IAD** [54] is a large UAD dataset recently released, containing 30 distinct objects. We follow the official splitting that includes all views, resulting in 36,465 normal images in the training set and 114,585 images in the test set (63,256 normal, 51,329 anomalous).

Metrics. Following prior works [17, 65], we adopt 7 evaluation metrics. Image-level 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 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 classes.

Implementation Details. ViT-Base/14 (patchsize=14) pre-trained by DINOv2-R [7] is used as the encoder by default. The drop rate of Noisy Bottleneck is 0.2 by default and increases to 0.4 on the diverse Real-IAD. Loose constraint with 2 groups is employed, and the anomaly map is given by the mean per-point cosine distance of the 2 groups. The input image is first resized to 448^2 and then center-cropped to 392^2, so the feature map (28^2) is large enough for anomaly localization. StableAdamW optimizer [56] with AMSGrad [41] (more stable than AdamW [35] in training) is utilized with lr=2e-3, β=(0.9,0.999) and wd=1e-4. The network is trained for 10,000 iterations (steps) on MVTec-AD and VisA, and 50,000 iterations on Real-IAD. Detailed settings are available in Appendix B.

4.2. Comparison to Multi-Class UAD SoTAs

We compare the proposed Dinomaly with the most advanced UAD and MUAD methods [10, 14, 17, 18, 34, 65, 67]. Experimental results are presented in Table 1, where Dinomaly surpasses compared methods by a large margin on all datasets and all metrics. On the most

表1. 多类别UAD设置下的性能（%）。†：专为MUAD设计的方法。Dinomaly↑：训练计划按比例扩大 到20,000、20,000和100,000次迭代（原始设置：10,000/10,000/50,000）。

Dateset	Method	Image-level			Pixel-level			
		AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
MVTec-AD [3]	RD4AD [10]	94.6	96.5	95.2	96.1	48.6	53.8	91.1
	SimpleNet [34]	95.3	98.4	95.8	96.9	45.9	49.7	86.5
	DeSTSeg [67]	89.2	95.5	91.6	93.1	54.3	50.9	64.8
	UniAD [60]†	96.5	98.8	96.2	96.8	43.4	49.5	90.7
	ReContrast [14]†	98.3	99.4	97.6	97.1	60.2	61.5	93.2
	DiAD [18]†	97.2	99.0	96.5	96.8	52.6	55.5	90.7
	ViTAD [65]†	98.3	99.4	97.3	97.7	55.3	58.7	91.4
	MambaAD [17]†	98.6	99.6	97.8	97.7	56.3	59.2	93.1
	Dinomaly (Ours)	**99.6**	**99.8**	**99.0**	**98.4**	**69.3**	**69.2**	**94.8**
	Dinomaly↑	99.7	99.8	99.2	98.4	69.3	69.2	94.7
VisA [70]	RD4AD [10]	92.4	92.4	89.6	98.1	38.0	42.6	91.8
	SimpleNet [34]	87.2	87.0	81.8	96.8	34.7	37.8	81.4
	DeSTSeg [67]	88.9	89.0	85.2	96.1	39.6	43.4	67.4
	UniAD [60]†	88.8	90.8	85.8	98.3	33.7	39.0	85.5
	ReContrast [14]†	95.5	96.4	92.0	98.5	47.9	50.6	91.9
	DiAD [18]†	86.8	88.3	85.1	96.0	26.1	33.0	75.2
	ViTAD [65]†	90.5	91.7	86.3	98.2	36.6	41.1	85.1
	MambaAD [17]†	94.3	94.5	89.4	98.5	39.4	44.0	91.0
	Dinomaly (Ours)	**98.7**	**98.9**	**96.2**	**98.7**	**53.2**	**55.7**	**94.5**
	Dinomaly↑	98.9	99.0	96.4	98.8	53.8	55.8	94.5
Real-IAD [54]	RD4AD [10]	82.4	79.0	73.9	97.3	25.0	32.7	89.6
	SimpleNet [34]	57.2	53.4	61.5	75.7	2.8	6.5	39.0
	DeSTSeg [67]	82.3	79.2	73.2	94.6	37.9	41.7	40.6
	UniAD [60]†	83.0	80.9	74.3	97.3	21.1	29.2	86.7
	ReContrast [14]†	86.4	84.2	77.4	97.8	31.6	38.2	91.8
	DiAD [18]†	75.6	66.4	69.9	88.0	2.9	7.1	58.1
	ViTAD [65]†	82.7	80.2	73.7	97.2	24.3	32.3	84.8
	MambaAD [17]†	86.3	84.6	77.0	98.5	33.0	38.7	90.5
	Dinomaly (Ours)	**89.3**	**86.8**	**80.2**	**98.8**	**42.8**	**47.1**	**93.9**
	Dinomaly↑	89.5	86.9	80.4	98.9	43.3	47.4	94.2

4. 实验

4.1. 实验设置

数据集。MVTec-AD [3] 包含15个类别（5个纹理类与10个物体类），共3,629张正常图像作为训练集，测试集包含1,725张图像（467张正常，1,258张异常）。VisA [70] 包含12个物体类别。训练集与测试集按官方划分方式拆分，得到训练集8,659张正常图像，测试集2,162张图像（962张正常，1,200张异常）。Real-IAD [54] 是近期发布的大规模无监督异常检测数据集，包含30个不同物体。我们遵循包含所有视角的官方划分方式，得到训练集36,465张正常图像，测试集114,585张图像（63,256张正常，51,329张异常）。

指标。遵循先前的研究[17, 65]，我们采用7种评估指标。图像级异常检测性能通过接收者操作特征曲线下面积（AUROC）、平均精度（AP）以及最优阈值下的F_1分数（F_1-max）来衡量。像素级异常定位通过AUROC、AP、F_1-max以及每区域重叠面积下面积（AUPRO）进行评估。结果数据集的平均是所有类别的平均值。

实现细节。默认使用DINOv2-R [7]预训练的ViT-Base/14（补丁尺寸=14）作为编码器。噪声瓶颈的丢弃率默认为0.2，在多样化的Real-IAD数据集上增至0.4。采用包含2组的宽松约束，异常图由两组间各点的平均余弦距离给出。输入图像首先调整尺寸至448^2，随后中心裁剪至392^2，从而确保特征图（28^2）足够大以支持异常定位。优化器采用结合AMSGrad [41]的StableAdamW [56]（训练中比AdamW [35]更稳定），参数设置为：学习率lr=2e-3，动量参数β=(0.9,0.999)，权重衰减wd=1e-4。网络在MVTec-AD和VisA数据集上训练10,000次迭代（步数），在Real-IAD数据集上训练50,000次迭代。详细配置见附录B。

4.2. 与多类别无监督异常检测先进方法的比较

我们将提出的Dinomaly与最先进的UAD和MUAD方法[10, 14, 17, 18, 34, 65, 67]进行了比较。实验结果如表1所示，其中Dinomaly在所有数据集和所有指标上都以显著优势超越了对比方法。在最具

Table 2. Performance under conventional **class-separated** UAD setting (%). n/a: not available.

Method	MVTec-AD [3]			VisA [70]			Real-IAD [54]		
	I-AUROC	P-AUROC	P-AUPRO	I-AUROC	P-AUROC	P-AUPRO	I-AUROC	P-AUROC	P-AUPRO
***Dinomaly** (MUAD)*	*99.6*	*98.4*	*94.8*	*98.7*	*98.7*	*94.5*	*89.3*	*98.8*	*93.9*
Dinomaly	**99.7**	**99.9**	**95.0**	**98.9**	**98.9**	**95.1**	**92.0**	**99.1**	**95.1**
RD4AD [10]	98.5	97.8	93.9	96.0	90.1	70.9	87.1	n/a	93.8
PatchCore [45]	99.1	98.1	93.5	94.7	98.5	91.8	89.4	n/a	91.5
SimpleNet [34]	99.6	98.1	90.0	97.1	98.2	90.7	88.5	n/a	84.6

Table 3. Ablations of Dinomaly elements on MVTec-AD (%). NB: Noisy Bottleneck. LA: Linear Attention. LC: Loose Constraint (2 groups). LL: Loose Loss. As MVTec-AD has reached saturation, we also present the results on VisA (Table A4).

NB	LA	LC	LL	Image-level			Pixel-level			
				AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
				98.41	99.09	97.41	97.18	62.96	63.82	92.95
✓				99.06	99.54	98.31	97.62	66.22	66.70	93.71
	✓			98.54	99.21	97.62	97.20	62.94	63.73	93.09
		✓		98.35	99.04	97.43	97.10	61.05	62.73	92.60
			✓	99.03	99.45	98.19	97.62	64.10	64.96	93.34
✓	✓			99.27	99.62	98.63	97.85	67.36	67.33	94.16
✓		✓		99.50	99.72	98.87	98.14	68.16	68.24	94.23
✓		✓	✓	99.52	99.73	98.92	98.20	68.25	68.34	94.17
✓	✓	✓		99.57	**99.78**	99.00	98.20	67.93	68.21	94.50
✓	✓	✓	✓	**99.60**	**99.78**	**99.04**	**98.35**	**69.29**	**69.17**	**94.79**

widely used MVTec-AD, Dinomaly produces image-level performance of **99.6/99.8/99.0** (%) and pixel-level performance of **98.4/69.3/69.2/94.8**, outperforming previous SoTAs by ***1.0/0.2/1.2*** and ***0.7/9.1/7.7/1.6***. This result declares that the image-level performance on the MVTec-AD dataset is nearly saturated under the MUAD setting. On the popular VisA, Dinomaly achieves image-level performance of **98.7/98.9/96.2** and pixel-level performance of **98.7/53.2/55.7/94.5**, outperforming previous SoTAs by ***3.2/2.5/4.2*** and ***0.2/5.3/5.1/2.6***. On the Real-IAD that contains 30 classes, each with 5 camera views, we produce image-level and pixel-level performance of **89.3/86.8/80.2** and **98.8/42.8/47.1/93.9**, outperforming previous SoTAs by ***3.0/2.2/3.2*** and ***0.3/4.9/5.4/3.4***, indicating our scalability to extremely complex scenarios. Per-class performances and qualitative visualization are presented in Appendix E and F. We also produce superior results on other popular UAD benchmarks, i.e., MPDD [24], BTAD [38], and Uni-Medical [66], with I-AUROC of 97.2, 95.4, and 84.9, respectively, as shown in Table A13 in Appendix.

4.3. Comparison to Class-Separated UAD SoTAs

Dinomaly is also compared with class-separated SoTAs, as shown in Table 2. Dinomaly under MUAD setting is comparable to conventional methods [10, 34, 45] that build individual models for each class. On MVTec-AD and VisA, multi-class Dinomaly (first row) is subjected to nearly no performance drop compared to its class-separated counterpart (second row). On the complicated Real-IAD that involves more classes and views, multi-class Dinomaly suffers a moderate performance drop but is still comparable to class-separated SoTAs.

4.4. Ablation Study

Overall Ablation. We conduct experiments to verify the effectiveness of the proposed elements, i.e., Noisy Bottleneck (NB), Linear Attention (LA), Loose Constraint (LC), and Loose Loss (LL). The already-powerful baseline (first row) is Dinomaly with noiseless MLP bottleneck, Softmax Attention, dense layer-to-layer supervision, and global cosine loss. This baseline is very similar to ViTAD [64] and the ViT version of RD4AD [10]. Results on MVTec-AD and VisA are shown in Table 3 and Table A4, respectively. NB and LL can directly contribute to the model performance. LA and LC boost the performance with the presence of NB. The use of LC is not solely beneficial because LC makes the reconstruction too easy without injected noise.

Model Scalability. Previous works [10, 60, 65] reported that anomaly detection networks do not follow the "scaling law". For example, RD4AD [10] found WideResNet50 better than WideResNet101 as the encoder backbone. ViTAD [65] found ViT-Small better than ViT-Base. On the contrary, as shown in Table 4, the performance of the proposed Dinomaly benefits from scaling. Dinomaly equipped with ViT-

表2. 常规类别分离UAD设置下的性能（%）。n/a：不可用。

Method	MVTec-AD [3]			VisA [70]			Real-IAD [54]		
	I-AUROC	P-AUROC	P-AUPRO	I-AUROC	P-AUROC	P-AUPRO	I-AUROC	P-AUROC	P-AUPRO
***Dinomaly** (MUAD)*	*99.6*	*98.4*	*94.8*	*98.7*	*98.7*	*94.5*	*89.3*	*98.8*	*93.9*
Dinomaly	**99.7**	**99.9**	**95.0**	**98.9**	**98.9**	**95.1**	**92.0**	**99.1**	**95.1**
RD4AD [10]	98.5	97.8	93.9	96.0	90.1	70.9	87.1	n/a	93.8
PatchCore [45]	99.1	98.1	93.5	94.7	98.5	91.8	89.4	n/a	91.5
SimpleNet [34]	99.6	98.1	90.0	97.1	98.2	90.7	88.5	n/a	84.6

表3. MVTec-AD上Dinomaly各模块的消融实验（%）。NB：噪声瓶颈。LA：线性注意力。LC：宽松约束（2组别）。LL：宽松损失。由于MVTec-AD已达到饱和，我们同时展示了VisA上的结果（表A4）。

NB	LA	LC	LL	Image-level			Pixel-level			
				AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
				98.41	99.09	97.41	97.18	62.96	63.82	92.95
✓				99.06	99.54	98.31	97.62	66.22	66.70	93.71
	✓			98.54	99.21	97.62	97.20	62.94	63.73	93.09
		✓		98.35	99.04	97.43	97.10	61.05	62.73	92.60
			✓	99.03	99.45	98.19	97.62	64.10	64.96	93.34
✓	✓			99.27	99.62	98.63	97.85	67.36	67.33	94.16
✓		✓		99.50	99.72	98.87	98.14	68.16	68.24	94.23
✓		✓	✓	99.52	99.73	98.92	98.20	68.25	68.34	94.17
✓	✓	✓		99.57	**99.78**	99.00	98.20	67.93	68.21	94.50
✓	✓	✓	✓	**99.60**	**99.78**	**99.04**	**98.35**	**69.29**	**69.17**	**94.79**

在广泛使用的MVTec-AD数据集上，Dinomaly实现了99.6/99.8/99.0 (%)的图像级性能以及98.4/69.3/69.2/94.8的像素级性能，分别以***1.0***/***0.2***/***1.2***和***0.7***/***9.1***/***7.7***/***1.6***的优势超越先前的最先进方法。这一结果表明，在MUAD设定下，MVTec-AD数据集的图像级性能已接近饱和。在流行的VisA数据集上，Dinomaly取得了98.7/98.9/96.2的图像级性能与98.7/53.2/55.7/94.5的像素级性能，分别以***3.2***/***2.5***/***4.2***和***0.2***/***5.3***/***5.1***/***2.6***的优势超越先前最优方法。在包含30个类别、每个类别具备5个相机视角的Real-IAD数据集上，我们实现了89.3/86.8/80.2的图像级性能与98.8/42.8/47.1/93.9的像素级性能，分别以***3.0***/***2.2***/***3.2***和***0.3***/***4.9***/***5.4***/***3.4***的优势超越先前最优方法，这证明了我们的方法对极端复杂场景的扩展能力。各类别详细性能与定性可视化结果见附录E和F。我们还在其他主流无监督异常检测基准（即MPDD [24]、BTAD [38]和UniMedical [66]）上取得了优异结果，其I-AUROC指标分别为97.2、95.4和84.9，详见附录中的表A13。

4.3. 与类别分离的无监督异常检测先进方法的比较

Dinomaly还与类别分离的SoTAs进行了比较，如表2所示。在MUAD设置下，Dinomaly与为每个类别建立独立模型的传统方法[10, 34, 45]表现相当。在MVTec-AD和VisA数据集上，多类别Dinomaly（首行）几乎未受到与按类别分离的对应方法（第二行）相比，性能有所下降。在涉及更多类别和视角的复杂Real-IAD数据集上，多类别Dinomaly虽然性能略有下降，但仍与按类别分离的当前最优方法（SoTAs）具有可比性。

4.4. 消融研究

总体消融实验。我们进行实验以验证所提出组件的有效性，即噪声瓶颈（NB）、线性注意力（LA）、宽松约束（LC）和宽松损失（LL）。已具备强大性能的基线（第一行）是采用无噪声MLP瓶颈、Softmax注意力、密集层间监督和全局余弦损失的Dinomaly。该基线与ViTAD [64] 及RD4AD [10] 的ViT版本非常相似。在MVTec-AD和VisA上的结果分别如表3和表A4所示。NB和LL可直接提升模型性能。LA和LC在NB存在时能进一步提升性能。单独使用LC并无益处，因为LC在没有注入噪声的情况下会使重建任务过于简单。

模型可扩展性。先前的研究[10, 60, 65]指出，异常检测网络并不遵循“缩放定律”。例如，RD4AD[10]发现WideResNet50作为编码器主干优于WideResNet101；ViTAD[65]发现ViT-Small优于ViT-Base。相反，如表4所示，所提出的Dino- maly性能受益于模型缩放。配备ViT-

Table 4. Scaling of ViT model sizes on MVTec-AD (%). Im/s (Troughoutput, image per second) is measured on NVIDIA RTX3090 with batch size=16. Results on VisA and Real-IAD are shown in Table A3. †:default.

Arch.	Params	MACs	Im/s	Image-level			Pixel-level			
				AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
ViT-Small	37.4M	26.3G	153.6	99.26	99.67	98.72	98.07	68.29	67.78	94.36
ViT-Base†	148.0M	104.7G	58.1	99.60	99.78	99.04	98.35	69.29	69.17	94.79
ViT-Large	275.3M	413.5G	24.2	**99.77**	**99.92**	**99.45**	**98.54**	**70.53**	**70.04**	**95.09**

Table 5. Scaling input size on MVTec-AD (%). †: default. Compared methods yield degradation when increasing input size.

Method	Input Size	Image-Level	Pixel-Level
RD4AD	256^2†	**94.6**/96.5/**96.1**	**96.1**/48.6/53.8/**91.1**
	320^2	93.2/**96.9**/95.6	95.7/**55.1/57.5/91.1**
	384^2	91.9/96.2/95.0	94.9/52.1/55.3/90.8
ReContrast	256^2†	**98.3/99.4/97.6**	**97.1**/60.2/61.5/93.2
	320^2	98.2/99.2/97.5	96.8/**61.8/62.6/93.3**
	384^2	95.2/98.0/96.4	96.5/57.7/59.5/92.6
Dinomaly	280^2	**99.6/99.8/99.3**	98.2/65.2/66.3/93.6
	336^2	**99.6/99.8**/99.2	98.3/67.2/67.8/94.2
	392^2†	**99.6/99.8**/99.0	**98.4/69.3/69.2/94.8**

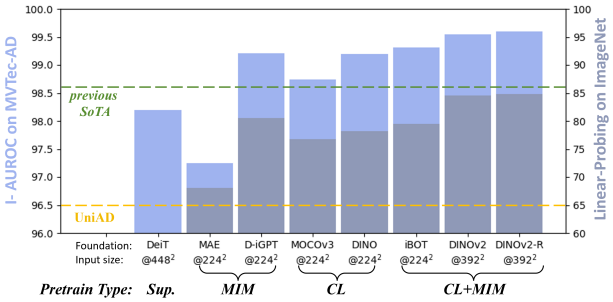


Figure 5. Image-level AUROC of Dinomaly equipped with various ViT foundations, and their linear-probing accuracy on ImageNet. MIM: Masked Image Modeling. CL: Contrastive Learning.

Small has already produced state-of-the-art results. ViT-Large further boosts Dinomaly to an unprecedented higher record. This scalability enables users to choose an appropriate model size based on the computational resources available in their specific scenario. A comparison of computational costs with other methods is presented in Table A11. In addition, training schedule can also be scaled up for even better performance without increasing inference costs, as demonstrated in Figure 1 (Dinomaly↑).

Input Scalability. Though it seems unfair to compare Dinomaly with previous works that take smaller images as input, we contend that increasing their input size not only fails to benefit but actively undermines their performance, especially for image-level detection performance, as shown in Table 5. Therefore, we follow the common comparison strategy based on "optimum vs. optimum". On the contrary, Dinomaly enjoys scaling input size for anomaly localization, while still producing SoTA performance given smaller images. Details are presented in Table A2 in Appendix.

ViT Foundations. We conduct extensive experiments to investigate the impact of diverse pre-trained ViT foundations, including DeiT [50], MAE [19], D-iGPT [44], MOCOv3 [6], DINO [4], iBot [69], DINOv2 [39], and DINOv2-R [7]. As shown in Figure 5, Dinomaly is robust to the choice of backbone. Almost all foundation models can produce SoTA-level results with image-level AUROC higher than 98%. The only notable exception is MAE, which, without fine-tuning, was reported to be less effective across various unsupervised tasks, e.g, kNN and linear-probing [39]. The optimal input size varies because the these backbones are pre-trained on different resolutions. Interestingly, we found the anomaly detection performance to be strongly correlated with the accuracy of ImageNet linear-probing (freeze backbone & only tune linear classifier) of the foundation model, suggesting the possibility of further improvement by simply adopting a future foundation model. Detailed results and analysis are presented in Appendix and Table A1.

Additional experiments and results are detailed in the Appendix C, encompassing evaluations of various pre-trained foundations, ablation studies of each components, hyperparameter optimization, and other in-depth analyses.

5. Conclusion

Dinomaly, a minimalistic UAD framework, is proposed to address the under-performed MUAD models in this paper. We present four key elements in Dinomaly, i.e., Foundation Transformer, Noisy MLP Bottleneck, Linear Attention, and Loose Reconstruction, that can boost the performance under the challenging MUAD setting without fancy modules and tricks. Extensive experiments on MVTec AD, VisA, and Real-IAD demonstrate our superiority over previous model-unified multi-class models and even recent class-separated models, indicating the feasibility of implementing a unified model in complicated scenarios free of severe performance degradation.

表4. MVTec-AD上ViT模型尺寸的缩放情况（%）。Im/s（吞吐量，每秒图像数）在NVIDIA RTX3090上测得 th
批处理大小=16。VisA和Real-IAD的结果如表A3所示。†：默认设置。

Arch.	Params	MACs	Im/s	Image-level			Pixel-level			
				AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
ViT-Small	37.4M	26.3G	153.6	99.26	99.67	98.72	98.07	68.29	67.78	94.36
ViT-Base†	148.0M	104.7G	58.1	99.60	99.78	99.04	98.35	69.29	69.17	94.79
ViT-Large	275.3M	413.5G	24.2	**99.77**	**99.92**	**99.45**	**98.54**	**70.53**	**70.04**	**95.09**

表5. MVTec-AD上输入尺寸的缩放效果（%）。†：默认设置。对比方法在增大输入尺寸时性能出现下降。

Method	Input Size	Image-Level	Pixel-Level
RD4AD	256^2†	**94.6**/96.5/**96.1**	**96.1**/48.6/53.8/**91.1**
	320^2	93.2/**96.9**/95.6	95.7/**55.1/57.5/91.1**
	384^2	91.9/96.2/95.0	94.9/52.1/55.3/90.8
ReContrast	256^2†	**98.3/99.4/97.6**	**97.1**/60.2/61.5/93.2
	320^2	98.2/99.2/97.5	96.8/**61.8/62.6/93.3**
	384^2	95.2/98.0/96.4	96.5/57.7/59.5/92.6
Dinomaly	280^2	**99.6/99.8/99.3**	98.2/65.2/66.3/93.6
	336^2	**99.6/99.8**/99.2	98.3/67.2/67.8/94.2
	392^2†	**99.6/99.8**/99.0	**98.4/69.3/69.2/94.8**

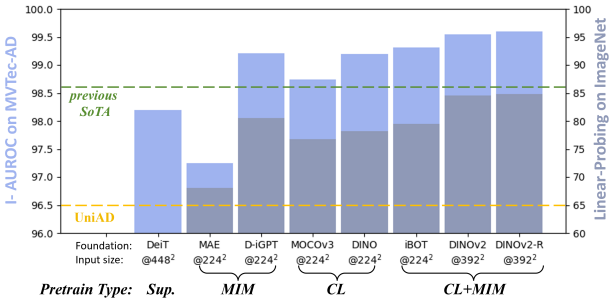


图5. 搭载不同ViT基座的Dinomaly在图像层面的AUROC性能，及其在ImageNet上的线性探测准确率。MIM：掩码图像建模。CL：对比学习。

小型模型已经取得了最先进的结果。ViT-大型模型进一步将Dinomaly提升至前所未有的更高记录。这种可扩展性使用户能够根据特定场景中可用的计算资源选择合适的模型规模。表A11展示了与其他方法计算成本的比较。此外，训练计划也可以进行扩展，以在不增加推理成本的情况下获得更好的性能，如图1（Dinomaly↑）所示。

输入可扩展性。尽管将Dinomaly与以往采用较小输入图像的工作进行比较似乎不公平，但我们认为增加其输入尺寸不仅无法带来益处，反而会损害其性能，尤其是在图像级检测性能方面，如表5所示。因此，我们遵循基于“最优对最优”的通用比较策略。相反，Dinomaly能够通过扩展输入尺寸来提升异常定位能力，同时在处理较小图像时仍能保持最先进的性能表现。详细数据见附录中的表A2。

ViT基础模型。我们进行了大量实验，以研究不同预训练ViT基础模型的影响，包括DeiT [50]、MAE [19]、D-iGPT [44]、MOCOv3 [6]、DINO [4]、iBot [69]、DINOv2 [39]和DINOv2-R [7]。如图5所示，Dinomaly对主干网络的选择具有鲁棒性。几乎所有基础模型都能产生SoTA级别的结果，图像级AUROC高于98%。唯一值得注意的例外是MAE，该模型未经微调时，在各种无监督任务（例如kNN和线性分类）中被报告为效果较差。

探测[39]。最优输入尺寸各不相同，因为这些骨干网络是在不同分辨率下预训练的。有趣的是，我们发现异常检测性能与基础模型在ImageNet线性探测（冻结骨干网络、仅调整线性分类器）的准确率高度相关，这表明通过直接采用未来的基础模型有可能实现进一步改进。详细结果与分析见附录及表A1。

额外的实验和结果详见附录C，包括对各种预训练基础的评估、各组成部分的消融研究、超参数优化以及其他深入分析。

5. 结论

本文提出了Dinomaly，一个极简的无监督异常检测框架，旨在解决当前多类别统一异常检测模型性能不足的问题。我们阐述了Dinomaly的四个核心要素：基础Transformer、带噪声的多层感知机瓶颈、线性注意力机制以及宽松重建策略。这些设计能够在极具挑战性的多类别统一异常检测设定下提升性能，且无需复杂的模块或技巧。在MVTec AD、VisA和Real-IAD数据集上的大量实验表明，我们的方法不仅优于以往的多类别统一模型，甚至超越了近期针对各类别单独训练的模型，这证明了在复杂场景中部署统一模型而不造成严重性能下降的可行性。

Acknowledgments

The authors acknowledge supports from National Natural Science Foundation of China (U22A2051, 82027807), National Key Research and Development Program of China (2022YFC2405200), Tsinghua-Foshan Innovation Special Fund (2021THFS0104), and Institute for Intelligent Healthcare, Tsinghua University (2022ZLB001).

References

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 622–637. Springer, 2019. 1

[2] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018. 1

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten 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*, pages 9592–9600, 2019. 1, 2, 6, 7

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3, 8, 2

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 3

[6] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 3, 8, 2

[7] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 3, 6, 8, 1, 2

[8] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 1

[9] David Dehaene and Pierre Eline. Anomaly localization by modeling perceptual features. *arXiv preprint arXiv:2008.05369*, 2020. 1

[10] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9737–9746, 2022. 5, 6, 7, 1, 8, 9, 10

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[13] Jia Guo, Shuai Lu, Lize Jia, Weihang Zhang, and Huiqi Li. Encoder-decoder contrast for unsupervised anomaly detection in medical images. *IEEE Transactions on Medical Imaging*, 2023. 1

[14] Jia Guo, Shuai Lu, Lize Jia, Weihang Zhang, and Huiqi Li. Recontrast: Domain-specific anomaly detection via contrastive reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10721–10740, 2023. 2, 3, 5, 6

[15] Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5961–5971, 2023. 5

[16] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. Diad: A diffusion-based framework for multi-class anomaly detection. *arXiv preprint arXiv:2312.06607*, 2023. 1, 3

[17] Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2404.06564*, 2024. 1, 3, 6, 7, 8, 10

[18] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8472–8480, 2024. 2, 6, 7, 8, 9, 10

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3, 8, 2, 5

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 5

[21] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. 4

[22] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *European Conference on Computer Vision*, pages 303–319. Springer, 2022. 7

[23] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. *arXiv preprint arXiv:2303.14814*, 2023. 7

[24] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of

致谢

作者感谢国家自然科学基金（U22A2051、82027807）、国家重点研发计划（2022YFC2405200）、清华大学-佛山创新专项基金（2021THFS0104）以及清华大学智能健康研究院（2022ZLB001）的支持。

参考文献

[1] Samet Akcay, Amir Atapour-Abarghouei 与 Toby P Breckon。Ganomaly：通过对抗训练进行半监督异常检测。收录于 *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*，第622–637页。Springer，2019年。1[2] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger 与 Carsten Steger。通过将结构相似性应用于自编码器来改进无监督缺陷分割。*arXiv preprint arXiv:1807.02011*，2018年。1[3] Paul Bergmann, Michael Fauser, David Sattlegger 与 Carsten Steger。Mvtec AD——一个用于无监督异常检测的综合性真实世界数据集。收录于 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*，第9592–9600页，2019年。1, 2, 6, 7[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski 与 Armand Joulin。自监督视觉Transformer中涌现的特性。收录于 *Proceedings of the IEEE/CVF international conference on computer vision*，第9650–9660页，2021年。3, 8, 2[5] Xinlei Chen 与 Kaiming He。探索简单的孪生网络表示学习。收录于 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*，第15750–15758页，2021年。3[6] Xinlei Chen, Saining Xie 与 Kaiming He。自监督视觉Transformer训练的实证研究。收录于 *Proceedings of the IEEE/CVF international conference on computer vision*，第9640–9649页，2021年。3, 8, 2[7] Timothée Darcet, Maxime Oquab, Julien Mairal 与 Piotr Bojanowski。视觉Transformer需要寄存器。*arXiv preprint arXiv:2309.16588*，2023年。3, 6, 8, 1, 2[8] Thomas Defard, Aleksandr Setkov, Angelique Loesch 与 Romaric Audigier。PaDiM：一种用于异常检测与定位的补丁分布建模框架。收录于 *International Conference on Pattern Recognition*，第475–489页。Springer，2021年。1[9] David Dehaene 与 Pierre Eline。通过建模感知特征进行异常定位。*arXiv preprint arXiv:2008.05369*，2020年。1

[[10] 邓寒秋，李星宇。基于单类嵌入反向蒸馏的异常检测。载于*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第9737–9746页，2022年。5, 6, 7, 1, 8, 9, 10

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: 一个大规模分层图像数据库。在 *IEEE Conference on Computer Vision and Pat- tern Recognition*，第248–255页，2009年。2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly 等。一幅图像值16x16个词：大规模图像识别的Transformer。*arXiv preprint arXiv:2010.11929*，2020年。2, 3[13] 郭佳，陆帅，贾立泽，张伟航，李慧琪。用于医学图像无监督异常检测的编码器-解码器对比。*IEEE Transactions on Medical Imaging*，2023年。1[14] 郭佳，陆帅，贾立泽，张伟航，李慧琪。Recontrast：通过对比重建实现领域特定的异常检测。于 *Advances in Neural Information Processing Systems (NeurIPS)*，第10721–10740页，2023年。2, 3, 5, 6[15] 韩东辰，潘旭然，韩一增，宋世绩，黄高。Flatten Transformer：使用聚焦线性注意力的视觉Transformer。于 *Proceedings of the IEEE/CVF International Conference on Computer Vision*，第5961–5971页，2023年。5[16] 何浩洋，张江宁，陈鸿旭，陈旭海，李志山，陈旭，王亚彪，王成杰，谢磊。DiAD：一个基于扩散的多类别异常检测框架。*arXiv preprint arXiv:2312.06607*，2023年。1, 3[17] 何浩洋，白宇虎，张江宁，何庆东，陈鸿旭，甘振业，王成杰，李相泰，田冠中，谢磊。MambaAD：探索用于多类别无监督异常检测的状态空间模型。*arXiv preprint arXiv:2404.06564*，2024年。1, 3, 6, 7, 8, 10[18] 何浩洋，张江宁，陈鸿旭，陈旭海，李志山，陈旭，王亚彪，王成杰，谢磊。一个基于扩散的多类别异常检测框架。于 *Proceedings of the AAAI Conference on Artifi- cial Intelligence*，第8472–8480页，2024年。2, 6, 7, 8, 9, 10[19] 何恺明，陈新雷，谢赛宁，李阳浩，Piotr Dollár，Ross Girshick。掩码自编码器是可扩展的视觉学习器。于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第16000–16009页，2022年。3, 8, 2, 5[20] Geoffrey Hinton, Oriol Vinyals, Jeff Dean。蒸馏神经网络中的知识。*arXiv preprint arXiv:1503.02531*，2015年。5[21] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, Ruslan R Salakhutdinov。通过防止特征检测器的共适应来改进神经网络。*arXiv preprint arXiv:1207.0580*，2012年。4[22] 黄超勤，管浩岩，姜傲凡，张亚，Michael Spratling，王延峰。基于配准的小样本异常检测。于 *European Conference on Computer Vision*，第303–319页。Springer，2022年。7[23] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, Onkar Dabeer。WinCLIP：零样本/少样本异常分类与分割。*arXiv preprint arXiv:2303.14814*，2023年。7[24] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, Milos Skotak。基于深度学习的缺陷检测

metal parts: evaluating current methods in complex conditions. In *2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pages 66–71. IEEE, 2021. 7, 6

[25] Xi Jiang, Jianlin Liu, Jinbao Wang, Qiang Nie, Kai Wu, Yong Liu, Chengjie Wang, and Feng Zheng. Softpatch: Unsupervised anomaly detection with noisy data. *Advances in Neural Information Processing Systems*, 35:15433–15445, 2022. 6, 7

[26] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020. 5

[27] Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10: 78446–78454, 2022. 1

[28] Yujin Lee, Harin Lim, and Hyunsoo Yoon. Selformaly: Towards task-agnostic unified anomaly detection. *arXiv preprint arXiv:2307.12540*, 2023. 3

[29] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021. 1

[30] Chen Liang, Jiahui Yu, Ming-Hsuan Yang, Matthew Brown, Yin Cui, Tuo Zhao, Boqing Gong, and Tianyi Zhou. Module-wise adaptive distillation for multimodality foundation models. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[31] Jiangqi Liu and Feng Wang. mixed attention auto encoder for multi-class industrial anomaly detection. *arXiv preprint arXiv:2309.12700*, 2023. 2

[32] Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J Radke, and Octavia Camps. Towards visually explaining variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8642–8651, 2020. 1

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[34] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. *arXiv preprint arXiv:2303.15140*, 2023. 6, 7, 1, 8, 9, 10

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[36] Ruiying Lu, YuJie Wu, Long Tian, Dongsheng Wang, Bo Chen, Xiyang Liu, and Ruimin Hu. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2310.14228*, 2023. 1, 2, 4

[37] Amira Ben Mabrouk and Ezzeddine Zagrouba. Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Systems with Applications*, 91:480–491, 2018. 1

[38] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE, 2021. 7, 6

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3, 8, 2

[40] 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. 3, 2

[41] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019. 6, 1

[42] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2806–2814, 2021. 7

[43] Tal Reiss, Niv Cohen, Eliahu Horwitz, Ron Abutbul, and Yedid Hoshen. Anomaly detection requires better representations. In *European Conference on Computer Vision*, pages 56–68. Springer, 2022. 3, 7

[44] Sucheng Ren, Zeyu Wang, Hongru Zhu, Junfei Xiao, Alan Yuille, and Cihang Xie. Rejuvenating image-gpt as strong visual representation learners. *arXiv preprint arXiv:2312.02147*, 2023. 8, 2

[45] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022. 7, 1, 3

[46] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912, 2021. 5, 1

[47] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019. 1

[48] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3531–3539, 2021. 5

[49] Shelly Sheynin, Sagie Benaim, and Lior Wolf. A hierarchical transformation-discriminating generative model for few shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8495–8504, 2021. 1

[50] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herv'e J'egou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2021. 8, 2

金属部件：在复杂条件下评估现有方法。于 *2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*，第66–71页。IEEE，2021年。7, 6 [25] 姜曦、刘建林、王金宝、聂强、吴凯、刘勇、王成杰、郑峰。Softpatch：使用噪声数据的无监督异常检测。*Advances in Neural Information Processing Systems*，35：15433–15445，2022年。6, 7 [26] Angelos Katharopoulos、Apoorv Vyas、Nikolaos Pappas、François Fleuret。Transformer是RNN：具有线性注意力的快速自回归Transformer。于*International conference on machine learning*，第5156–5165页。PMLR，2020年。5 [27] 李成旭、李承贤、宋炳哲。CFA：基于耦合超球面的特征自适应用于目标导向的异常定位。*IEEE Access*，10：78446–78454，2022年。1 [28] 李裕珍、林河仁、尹贤秀。Selformaly：迈向任务无关的统一异常检测。*arXiv preprint arXiv:2307.12540*，2023年。3 [29] 李春亮、孙基赫、尹镇成、Tomas Pfister。CutPaste：用于异常检测与定位的自监督学习。于*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第9664–9674页，2021年。1 [30] 梁晨、俞嘉慧、杨明轩、Matthew Brown、崔寅、赵拓、龚伯清、周天一。模块化自适应蒸馏用于多模态基础模型。*Advances in Neural Information Processing Systems*，36，2024年。5 [31] 刘江奇、王峰。用于多类工业异常检测的混合注意力自编码器。*arXiv preprintarXiv:2309.12700*，2023年。2 [32] 刘文倩、李润泽、郑萌、Srikrishna Karanam、吴子彦、Bir Bhanu、Richard J Radke、Octavia Camps。面向视觉解释变分自编码器。于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第8642–8651页，2020年。1 [33] 刘泽、林宇彤、曹越、胡涵、魏亦轩、张峥、林史蒂芬、郭百宁。Swin Transformer：使用移位窗口的分层视觉Transformer。于 *IEEE/CVF International Conference on Computer Vision*，第10012–10022页，2021年。3 [34] 刘志康、周益铭、徐元生、王子雷。Simplenet：用于图像异常检测与定位的简单网络。*arXiv preprint arXiv:2303.15140*，2023年。6, 7, 1, 8, 9, 10 [35] Ilya Loshchilov、Frank Hutter。解耦权重衰减正则化。*arXiv preprint arXiv:1711.05101*，2017年。6 [36] 卢瑞英、吴宇杰、田龙、王东升、陈波、刘西洋、胡瑞敏。用于多类无监督异常检测的分层向量量化Transformer。*arXiv preprint arXiv:2310.14228*，2023年。1, 2, 4 [37] Amira Ben Mabrouk、Ezzeddine Zagrouba。智能视频监控系统中的异常行为识别：综述。*Expert Systems with Applications*，91：480–491，2018年。1

[38] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. VT-ADL：一种用于图像异常检测与定位的视觉Transformer网络。发表于 *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*，第01–06页。IEEE，2021年。7, 6[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, 等。Dinov2：无需监督学习鲁棒的视觉特征。*arXiv preprint arXiv:2304.07193*，2023年。3, 8, 2[40] Z Peng, L Dong, H Bao, Q Ye, and F Wei。BEiT v2：使用向量量化视觉标记器的掩码图像建模。*arXiv preprint arXiv:2208.06366*，2022年。3, 2[41] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar。论Adam及其后续算法的收敛性。*arXiv preprint arXiv:1904.09237*，2019年。6, 1[42] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen。PANDA：为异常检测与分割适配预训练特征。发表于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第2806–2814页，2021年。7[43] Tal Reiss, Niv Cohen, Eliahu Horwitz, Ron Abutbul, and Yedid Hoshen。异常检测需要更好的表征。发表于 *European Conference on Computer Vision*，第56–68页。Springer，2022年。3, 7[44] Sucheng Ren, Zeyu Wang, Hongru Zhu, Junfei Xiao, Alan Yuille, and Cihang Xie。重振Image-GPT作为强大的视觉表征学习器。*arXiv preprint arXiv:2312.02147*，2023年。8, 2[45] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler。迈向工业异常检测的完全召回。发表于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第14318–14328页，2022年。7, 1, 3[46] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee。用于异常检测的多分辨率知识蒸馏。发表于 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*，第14902–14912页，2021年。5, 1[47] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth。f-AnoGAN：基于生成对抗网络的快速无监督异常检测。*Medical image analysis*，54卷：30–44页，2019年。1[48] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li。高效注意力：具有线性复杂度的注意力机制。发表于 *Proceedings of the IEEE/CVF winter conference on applications of computer vision*，第3531–3539页，2021年。5[49] Shelly Sheynin, Sagie Benaim, and Lior Wolf。一种用于少样本异常检测的分层变换判别生成模型。发表于 *Proceedings of the IEEE/CVF International Conference on Computer Vision*，第8495–8504页，2021年。1[50] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herv’e J’egou。通过注意力训练数据高效的图像Transformer并进行蒸馏。*arXiv preprint arXiv:2012.12877*，2021年。8, 2

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 2

[52] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 4

[53] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. In *Journal of Machine Learning Research*, pages 3371–3408, 2010. 4

[54] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jianning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. *arXiv preprint arXiv:2403.12580*, 2024. 2, 6, 7, 3

[55] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for anomaly detection. In *The British Machine Vision Conference (BMVC)*, 2021. 1

[56] Mitchell Wortsman, Tim Dettmers, Luke Zettlemoyer, Ari Morcos, Ali Farhadi, and Ludwig Schmidt. Stable and low-precision training for large-scale vision-language models. *Advances in Neural Information Processing Systems*, 36: 10271–10298, 2023. 6, 1

[57] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022. 3

[58] Jie Yang, Yong Shi, and Zhiquan Qi. Dfr: Deep feature reconstruction for unsupervised anomaly segmentation. *arXiv preprint arXiv:2012.07122*, 2020. 5, 1

[59] Haonan Yin, Guanlong Jiao, Qianhui Wu, Borje F Karlsson, Biqing Huang, and Chin Yew Lin. Lafite: Latent diffusion model with feature editing for unsupervised multi-class anomaly detection. *arXiv preprint arXiv:2307.08059*, 2023. 1, 2, 3, 4

[60] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *arXiv preprint arXiv:2206.03687*, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

[61] Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. Metaformer baselines for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 4

[62] Vitjan Zavrtanik, Matej Kristan, and Danijel Skoč aj. Draema discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021. 4, 1

[63] Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706, 2021. 1

[64] Dingwen Zhang, Guohai Huang, Qiang Zhang, Jungong Han, Junwei Han, Yizhou Wang, and Yizhou Yu. Exploring task structure for brain tumor segmentation from multimodality mr images. *IEEE Transactions on Image Processing*, 29:9032–9043, 2020. 7

[65] Jiangning Zhang, Xuhai Chen, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, Ming-Hsuan Yang, and Dacheng Tao. Exploring plain vit reconstruction for multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2312.07495*, 2023. 3, 5, 6, 7

[66] Jiangning Zhang, Haoyang He, Zhenye Gan, Qingdong He, Yuxuan Cai, Zhucun Xue, Yabiao Wang, Chengjie Wang, Lei Xie, and Yong Liu. Ader: A comprehensive benchmark for multi-class visual anomaly detection. *arXiv preprint arXiv:2406.03262*, 2024. 7, 2, 6

[67] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3914–3923, 2023. 4, 6, 1, 7, 8, 9, 10

[68] Ying Zhao. Omnial: A unified cnn framework for unsupervised anomaly localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3924–3933, 2023. 1, 3

[69] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 3, 8, 2

[70] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022. 2, 6, 7, 3

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, Illia Polosukhin。注意力就是您所需要的一切。发表于 *Advances in Neural Information Processing Systems*，第 5998–6008 页，2017 年。2[52] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, Pierre-Antoine Manzagol。使用去噪自编码器提取和组合鲁棒特征。发表于 *Proceedings of the 25th international conference on Machine learning*，第 1096–1103 页，2008 年。4[53] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol。堆叠去噪自编码器：在具有局部去噪准则的深度网络中学习有用的表示。发表于 *Journal of Machine Learning Research*，第 3371–3408 页，2010 年。4[54] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jianning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, Lizhuang Ma。Real-IAD：一个用于基准测试多功能工业异常检测的真实世界多视角数据集。*arXiv preprint arXiv:2403.12580*，2024 年。2, 6, 7, 3[55] Guodong Wang, Shumin Han, Errui Ding, Di Huang。用于异常检测的学生-教师特征金字塔匹配。发表于 *The British Machine Vision Conference (BMVC)*，2021 年。1[56] Mitchell Wortsman, Tim Dettmers, Luke Zettlemoyer, Ari Morcos, Ali Farhadi, Ludwig Schmidt。大规模视觉语言模型的稳定低精度训练。*Advances in Neural Information Processing Systems*，第 36 卷：10271–10298，2023 年。6, 1[57] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, Han Hu。SimMIM：一个用于掩码图像建模的简单框架。发表于 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*，第 9653–9663 页，2022 年。3[58] Jie Yang, Yong Shi, Zhiquan Qi。DFR：用于无监督异常分割的深度特征重建。*arXiv preprint arXiv:2012.07122*，2020 年。5, 1[59] Haonan Yin, Guanlong Jiao, Qianhui Wu, Borje F Karlsson, Biqing Huang, Chin Yew Lin。LaFite：具有特征编辑的潜在扩散模型用于无监督多类异常检测。*arXiv preprint arXiv:2307.08059*，2023 年。1, 2, 3, 4[60] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, Xinyi Le。一个用于多类异常检测的统一模型。*arXiv preprint arXiv:2206.03687*，2022 年。1, 2, 3, 4, 5, 6, 7, 8, 9, 10[61] Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, Xinchao Wang。视觉任务的 Metaformer 基线。*IEEE Transactions on Pattern Analysis and Machine Intelligence*，2023 年。4[62] Vitjan Zavrtanik, Matej Kristan, Danijel Skočaj。DRAEM - 一种用于表面异常检测的判别性训练重建嵌入。发表于 *Proceedings of the IEEE/CVF International Conference on Computer Vision*，第 8330–8339 页，2021 年。4, 1[63] Vitjan Zavrtanik, Matej Kristan, Danijel Skocaj。通过修复进行重建的视觉异常检测。*Pattern Recognition*，第 112 卷：107706，2021 年。1[64] 张定文，黄国海，张强，韩军功，韩军伟，王一舟，俞益州。探索多模态磁共振图像中脑肿瘤分割的任务结构。*IEEE Transactions on Image Process- ing*，第29卷，第9032–9043页，2020年。7[65] 张江宁，陈旭海，王亚彪，王成杰，刘勇，李相泰，杨明炫，陶大程。探索朴素视觉Transformer重建用于多类别无监督异常检测。*arXiv preprint arXiv:2312.07495*，2023年。3, 5, 6, 7[66] 张江宁，何浩洋，甘振业，何庆东，蔡宇轩，薛朱存，王亚彪，王成杰，谢磊，刘勇。ADER：一个全面的多类别视觉异常检测基准。*arXiv preprint arXiv:2406.03262*，2024年。7, 2, 6[67] 张璇，李世宇，李曦，黄平，单九龙，陈婷。DestSeg：用于异常检测的分割引导去噪师生模型。收录于*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第3914–3923页，2023年。4, 6, 1, 7, 8, 9, 10[68] 赵颖。OmniAL：一个用于无监督异常定位的统一CNN框架。收录于*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第3924–3933页，2023年。1, 3[69] 周靖皓，魏晨，王慧宇，沈伟，谢慈航，Alan Yuille，孔涛。iBOT：使用在线分词器的图像BERT预训练。*arXiv preprint arXiv:2111.07832*，2021年。3, 8, 2[70] 邹阳，Jongheon Jeong，Latha Pemula，张东清，Onkar Dabeer。用于异常检测与分割的"找不同"自监督预训练。收录于*Eu- ropean Conference on Computer Vision*，第392–408页。Springer出版社，2022年。2, 6, 7, 3

Dinomaly: The *Less Is More* Philosophy in Multi-Class Unsupervised Anomaly Detection

Supplementary Material

A. Additional Related Work

Here, we discussed general methods for unsupervised anomlay detection. *Epistemic methods* are based on the assumption that the networks respond differently during inference between seen input and unseen input. Within this paradigm, *pixel reconstruction* methods assume that the networks trained on normal images can reconstruct anomaly-free regions well, but poorly for anomalous regions. Auto-encoder (AE) [2, 63], variational auto-encoder (VAE) [9, 32], or generative adversarial network (GAN) [1, 47] are used to restore normal pixels. However, *pixel reconstruction* models may also succeed in restoring unseen anomalous regions if they resemble normal regions in pixel values or the anomalies are barely noticeable [10]. Therefore, *feature reconstruction* is proposed to construct features of pre-trained encoders instead of raw pixels [10, 58, 60]. To prevent the whole network from converging to a trivial solution, the parameters of the encoders are frozen during training. In *feature distillation* [46, 55], the student network is trained from scratch to mimic the output features of the pre-trained teacher network with the same input of normal images, also based on the similar hypothesis that the student trained on normal samples only succeed in mimicking features of normal regions.

Pseudo-anomaly methods generate handcrafted defects on normal images to imitate anomalies, converting UAD to supervised classification [29] or segmentation tasks [62]. Specifically, CutPaste [29] simulates anomalous regions by randomly pasting cropped patches of normal images. DRAEM [62] constructs abnormal regions using Perlin noise as the mask and another image as the additive anomaly. DeTSeg [67] employs a similar anomaly generation strategy and combines it with feature reconstruction. SimpleNet [34] introduces anomaly by injecting Gaussian noise in the pre-trained feature space. These methods deeply rely on how well the pseudo anomalies match the real anomalies, which makes it hard to generalize to different datasets.

Feature statistics methods [8, 27, 45, 49] 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. 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.

Scope of Application. In this work, we focus on **sensory AD** that detects regional or structural anomalies (common in practical applications such as industrial inspection, medical disease screening, etc.), which is distinguished from **semantic AD**. In sensory AD, normal and anomalous samples are the same objects except for anomaly, e.g. good cable vs. spoiled cable. In semantic AD, the class of normal samples and anomalous samples are semantically different, e.g. animals vs. vehicles. Semantic AD methods usually utilize and compare the global representation of images, which generally do not suffer from the issues of multi-class setting discussed in this paper..

B. Full Implementation Details

ViT-Base/14 (patch size=14) pre-trained by DINOv2 with registers (DINOv2-R) [7] is utilized as the encoder by default. The discard rate of Dropout in Noisy Bottleneck is 0.2 by default, which is increased to 0.4 for the diverse Real-IAD. Loose constraint with 2 groups and $\mathcal{L}_{global-hm}$ loss are used by default. The input image is first resized to 448^2 and then center-cropped to 392^2, so that the feature map (28^2) is large enough for localization. As previously discussed, the middle 8 layers of 12-layer ViT-Base are used for reconstruction and feeding the bottleneck. ViT-Small also has 12 layers, which is the same. ViT-Large contains 24 layers; therefore, we use the [4,6,8,...18] layers (index start from 0). The decoder always contains 8 layer.

StableAdamW optimizer [56] with AMSGrad [41] is utilized with lr (learning rate)=2e-3, β=(0.9,0.999), wd (weight decay)=1e-4 and eps=1e-10. The network is trained for 10,000 iterations for MVTec-AD and VisA and 50,000 iterations for Real-IAD under MUAD setting. The network is trained for 5,000 iterations on each class under the class-separated UAD setting. The lr warms up from 0 to 2e-3 in the first 100 iterations and cosine anneals to 2e-4 throughout the training. The discarding rate in Equation 5 linearly rises from 0% to 90% in the first 1,000 iterations as warm-up (500 iters for class-separated setting). The anomaly map is obtained by upsampling the point-wise cosine distance between encoder and decoder feature maps (averaging if more than one pair or group). The mean of the top 1% pixels in an anomaly map is used as the image anomaly score. All experiments are conducted with random seed=1 with cuda deterministic for invariable weight initialization and batch order. Codes are implemented with Python 3.8 and PyTorch 1.12.0 cuda 11.3, and run on NVIDIA GeForce RTX3090 GPUs (24GB).

Most results of compared MUAD SoTAs are directly

Dinomaly：多类无监督异常检测中的 *Less Is More*哲学

补充材料

A. 其他相关工作

在此，我们讨论了无监督异常检测的通用方法。*Epistemic methods*基于这样的假设：网络在推理过程中对已见输入和未见输入的反应不同。在这一范式下，*pixel reconstruction*方法假设在正常图像上训练的网络能够很好地重建无异常区域，但对异常区域的重建效果较差。自编码器（AE）[2, 63]、变分自编码器（VAE）[9, 32]或生成对抗网络（GAN）[1, 47]被用于恢复正常像素。然而，*pixel reconstruction*模型也可能成功恢复未见过的异常区域，如果这些区域在像素值上与正常区域相似，或者异常几乎难以察觉[10]。因此，*feature reconstruction*被提出用于构建预训练编码器的特征，而非原始像素[10, 58, 60]。为防止整个网络收敛到平凡解，编码器的参数在训练期间被冻结。在*feature distillation*[46, 55]中，学生网络从零开始训练，以模仿预训练教师网络在相同正常图像输入下的输出特征，该方法同样基于类似的假设：仅在正常样本上训练的学生网络只能成功模仿正常区域的特征。

Pseudo-anomaly 方法通过在正常图像上生成手工制作的缺陷来模拟异常，将无监督异常检测（UAD）转化为有监督的分类[29]或分割任务[62]。具体而言，CutPaste[29]通过随机粘贴正常图像的裁剪区域来模拟异常区域。DRAEM[62]使用Perlin噪声作为掩码，并结合另一张图像作为叠加异常来构建异常区域。DeTSeg[67]采用了类似的异常生成策略，并将其与特征重建相结合。SimpleNet[34]通过在预训练特征空间中注入高斯噪声来引入异常。这些方法高度依赖于伪异常与真实异常的匹配程度，因此难以泛化到不同的数据集。

Feature statistics 方法[8, 27, 45, 49]会记忆所有通过在大规模数据集上预训练的网络提取的正常特征（或其建模分布），并在推理过程中将其与测试样本进行匹配。由于这些方法需要记忆、处理和匹配来自训练样本的几乎所有特征，它们在训练和推理阶段都计算成本高昂，尤其是在训练集规模较大时。

适用范围。在本研究中，我们主要关注感官检测区域或结构异常（常见于工业检测、医学疾病筛查等实际应用）的异常检测，与语义异常检测有所区别。在感官异常检测中，正常样本与异常样本除异常外属于同一类对象，例如完好的电缆与损坏的电缆。而在语义异常检测中，正常样本与异常样本的类别在语义上不同，例如动物与车辆。语义异常检测方法通常利用并比较图像的全局表示，这类方法通常不受本文讨论的多类别设置问题的影响。

B. 完整实现细节

默认采用DINOv2带寄存器版本（DINOv2-R）[7]预训练的ViT-Base/14（图像块尺寸=14）作为编码器。噪声瓶颈中Dropout的默认丢弃率为0.2，在多样化的Real-IAD任务中提升至0.4。默认使用2组松散约束与$\mathcal{L}_{global-hm}$损失函数。输入图像首先缩放至448^2像素，随后中心裁剪为392^2像素，以确保生成的特征图（28^2）具备足够分辨率进行定位。如前所述，12层ViT-Base的中间8层被用于重建并馈入瓶颈层。ViT-Small同样为12层结构，采用相同配置。ViT-Large包含24层，因此采用[4,6,8,...18]层（索引从0开始）。解码器始终由8层构成。

采用StableAdamW优化器[56]结合AMSGrad[41]，设置学习率为2e-3，β参数为(0.9, 0.999)，权重衰减为1e-4，ε为1e-10。网络在MVTec-AD和VisA数据集上训练10,000次迭代，在Real-IAD数据集的MUAD设置下训练50,000次迭代；在按类别分离的UAD设置中，每个类别训练5,000次迭代。学习率在前100次迭代中从0预热至2e-3，随后通过余弦退火逐渐下降至2e-4。公式5中的丢弃率在前1,000次迭代中从0%线性上升至90%作为预热（在类别分离设置中为500次迭代）。异常图通过对编码器与解码器特征图之间的逐点余弦距离进行上采样获得（若存在多对或多组特征图则取平均）。异常图中前1%像素的均值被用作图像异常分数。所有实验均采用随机种子1，并启用CUDA确定性以保证权重初始化和批次顺序不变。代码基于Python 3.8和PyTorch 1.12.0（CUDA 11.3）实现，运行于NVIDIA GeForce RTX3090 GPU（24GB显存）。

相比MUAD SoTAs的大多数结果直接

Table A1. Comparison between pre-trained ViT foundations, conducted on MVTec-AD (%). All models are ViT-Base. The patch size of DINOv2 and DINOv2-R is 14^2; others are 16^2. R448^2-C392^2 represents first resizing images to 448×448, then center cropping to 392×392.

Pre-Train Backbone	Type	Image Size	Image-level			Pixel-level			
			AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
DeiT[50]	Supervised	R512^2-C448^2	98.19	99.24	97.64	97.93	68.98	67.91	91.45
MAE[19]	MIM	R512^2-C448^2	96.27	98.33	95.44	96.96	62.89	63.32	89.85
D-iGPT[44]	MIM	R512^2-C448^2	98.75	99.24	97.70	98.30	65.77	66.16	92.34
MOCOv3[6]	CL	R512^2-C448^2	98.47	99.42	97.36	98.52	70.99	69.41	92.83
DINO[4]	CL	R512^2-C448^2	98.97	99.58	98.14	98.52	70.89	69.02	93.48
iBOT[69]	CL+MIM	R512^2-C448^2	99.22	99.67	98.57	98.60	70.78	69.92	93.33
DINOv2[39]	CL+MIM	R448^2-C392^2	99.55	99.81	99.13	98.26	68.35	68.79	94.83
DINOv2-R[7]	CL+MIM	R448^2-C392^2	99.60	99.78	99.04	98.35	69.29	69.17	94.79
DeiT[50]	Supervised	R256^2-C224^2	97.65	99.05	97.40	97.80	62.58	63.39	89.98
MAE[19]	MIM	R256^2-C224^2	97.25	98.84	96.94	97.78	63.00	64.01	90.95
BEiTv2[40]	MIM	R256^2-C224^2	97.70	99.11	97.39	97.61	59.79	62.53	90.10
D-iGPT[44]	MIM	R256^2-C224^2	99.21	99.66	98.47	98.08	60.05	63.05	91.78
MOCOv3[6]	CL	R256^2-C224^2	98.74	99.56	98.33	98.05	63.36	64.38	91.13
DINO[4]	CL	R256^2-C224^2	99.20	99.72	98.77	98.16	64.16	65.07	92.02
iBOT[69]	CL+MIM	R256^2-C224^2	99.31	99.74	98.77	98.25	64.01	65.37	91.68
DINOv2[39]	CL+MIM	R256^2-C224^2	99.26	99.70	98.60	97.95	62.27	64.39	92.80
DINOv2-R[7]	CL+MIM	R256^2-C224^2	99.34	99.73	99.03	98.09	63.04	64.48	92.59

Table A2. Ablations of input size, conducted on MVTec-AD (%). R448^2-C392^2 represents first resizing images to 448×448, then center cropping to 392×392.

Image Size	MACs	Image-level			Pixel-level			
		AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
R512^2-C448^2	136.4G	**99.67**	99.81	99.12	98.33	69.24	**69.47**	94.76
R448^2	136.4G	99.59	99.77	99.19	**98.57**	68.09	68.58	**95.60**
R448^2-C392^2†	104.7G	99.60	99.78	99.04	98.35	**69.29**	**69.17**	94.79
R392^2	104.7G	99.48	99.74	99.04	98.47	67.02	67.86	95.34
R384^2-C336^2	77.1G	99.61	99.78	99.22	98.27	67.22	67.77	94.24
R336^2	77.1G	99.63	**99.84**	99.23	98.48	65.46	66.60	95.10
R320^2-C280^2	53.7G	99.62	99.81	99.07	98.21	65.21	66.34	93.57
R280^2	53.7G	99.46	99.75	**99.27**	98.40	63.28	64.79	94.47

drawn from a benchmark paper ADer [66]. We express great thanks for their wonderful work.

C. Additional Ablation and Experiment

Pre-Trained Foundations. The representation quality of the frozen backbone Transformer is of great significance to unsupervised anomaly detection. We conduct extensive experiments to probe the impact of different pre-training methods, including supervised learning and self-supervised learning. DeiT [50] is trained on ImageNet[11] in a supervised manner by distilling CNNs. MAE [19], BEiTv2 [40], and D-iGPT [44] are based on masked image modeling (MIM). Given input images with masked patches, MAE [19] is optimized to restore raw pixels; BEiTv2 [40] is trained to predict the token index of VQ-GAN and CLIP; D-iGPT [44] is trained to predict the features of CLIP model. MOCOv3 [6] is based on contrastive learning (CL), pulling the representations of the similar images and pushing those of different images. DINO [4] is based on positive-pair contrastive learning, which is also referred to as self-distillation. It trains the network to produce similar feature representations given two views (augmentations) of the same image. iBot [69] and DINOv2 [39] combine MIM and CL strategies, marking the SoTA of self-supervised foundation models. DINOv2-R [7] is a variation

表 A1. 在 MVTEC-AD (%) 上进行的预训练 ViT 基础模型对比。所有模型均为 ViT-Base 架构。DINOv2 与 DINOv2-R 的补丁尺寸为 14^2；其他模型为 16^2。R448^2-C392^2 表示先将图像缩放至 448××448，再中心裁剪至 392××392。

Pre-Train Backbone	Type	Image Size	Image-level			Pixel-level			
			AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
DeiT[50]	Supervised	R512^2-C448^2	98.19	99.24	97.64	97.93	68.98	67.91	91.45
MAE[19]	MIM	R512^2-C448^2	96.27	98.33	95.44	96.96	62.89	63.32	89.85
D-iGPT[44]	MIM	R512^2-C448^2	98.75	99.24	97.70	98.30	65.77	66.16	92.34
MOCOv3[6]	CL	R512^2-C448^2	98.47	99.42	97.36	98.52	70.99	69.41	92.83
DINO[4]	CL	R512^2-C448^2	98.97	99.58	98.14	98.52	70.89	69.02	93.48
iBOT[69]	CL+MIM	R512^2-C448^2	99.22	99.67	98.57	98.60	70.78	69.92	93.33
DINOv2[39]	CL+MIM	R448^2-C392^2	99.55	99.81	99.13	98.26	68.35	68.79	94.83
DINOv2-R[7]	CL+MIM	R448^2-C392^2	99.60	99.78	99.04	98.35	69.29	69.17	94.79
DeiT[50]	Supervised	R256^2-C224^2	97.65	99.05	97.40	97.80	62.58	63.39	89.98
MAE[19]	MIM	R256^2-C224^2	97.25	98.84	96.94	97.78	63.00	64.01	90.95
BEiTv2[40]	MIM	R256^2-C224^2	97.70	99.11	97.39	97.61	59.79	62.53	90.10
D-iGPT[44]	MIM	R256^2-C224^2	99.21	99.66	98.47	98.08	60.05	63.05	91.78
MOCOv3[6]	CL	R256^2-C224^2	98.74	99.56	98.33	98.05	63.36	64.38	91.13
DINO[4]	CL	R256^2-C224^2	99.20	99.72	98.77	98.16	64.16	65.07	92.02
iBOT[69]	CL+MIM	R256^2-C224^2	99.31	99.74	98.77	98.25	64.01	65.37	91.68
DINOv2[39]	CL+MIM	R256^2-C224^2	99.26	99.70	98.60	97.95	62.27	64.39	92.80
DINOv2-R[7]	CL+MIM	R256^2-C224^2	99.34	99.73	99.03	98.09	63.04	64.48	92.59

表A2。在MVTec-AD上进行的输入尺寸消融实验（%）。R448^2-C392^{2}表示先将图像尺寸调整为448××448，再进行中心裁剪至392××392。

Image Size	MACs	Image-level			Pixel-level			
		AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
R512^2-C448^2	136.4G	**99.67**	99.81	99.12	98.33	<u>69.24</u>	**69.47**	94.76
R448^2	136.4G	99.59	99.77	99.19	**98.57**	68.09	68.58	**95.60**
R448^2-C392^2†	104.7G	99.60	99.78	99.04	98.35	**69.29**	**69.17**	94.79
R392^2	104.7G	99.48	99.74	99.04	98.47	67.02	67.86	<u>95.34</u>
R384^2-C336^2	77.1G	99.61	99.78	99.22	98.27	67.22	67.77	94.24
R336^2	77.1G	<u>99.63</u>	**99.84**	<u>99.23</u>	<u>98.48</u>	65.46	66.60	95.10
R320^2-C280^2	53.7G	99.62	<u>99.81</u>	99.07	98.21	65.21	66.34	93.57
R280^2	53.7G	99.46	99.75	**99.27**	98.40	63.28	64.79	94.47

摘自基准论文ADer [66]。我们对他们的出色工作表示衷心感谢。

C. 额外的消融实验与验证

预训练基础。冻结骨干Transformer的表示质量对无监督异常检测至关重要。我们进行了大量实验来探究不同预训练方法的影响，包括监督学习和自监督学习。DeiT[50]通过蒸馏CNN在ImageNet[11]上进行监督式训练。MAE[19]、BEiTv2[40]和D-iGPT[44]则基于掩码图年龄建模（MIM）。给定带有掩码补丁的输入图像，MAE [19] 被优化以恢复原始像素；BEiTv2 [40] 被训练用于预测 VQ-GAN 和 CLIP 的标记索引；D-iGPT [44] 被训练用于预测 CLIP 模型的特征。MOCOv3 [6] 基于对比学习（CL），拉近相似图像的表示并推远不同图像的表示。DINO [4] 基于正样本对对比学习，也被称为自蒸馏。它训练网络对同一图像的两个视图（增强版本）产生相似的特征表示。iBot [69] 和 DINOv2 [39] 结合了 MIM 和 CL 策略，标志着自监督基础模型的当前最佳水平。DINOv2-R [7] 是一个变体。

Table A3. Scaling of ViT architectures on VisA and Real-IAD (%). †: default.

Dateset	Arch.	Image-level			Pixel-level			
		AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
VisA [70]	ViT-Small	97.94	98.09	95.33	98.57	51.19	55.10	93.71
	ViT-Base†	98.73	98.87	**96.18**	98.74	53.23	55.69	94.50
	ViT-Large	**98.85**	**99.09**	96.12	**99.10**	**55.68**	**57.33**	**94.76**
Real-IAD [54]	ViT-Small	89.10	86.91	79.87	98.69	41.88	46.74	94.08
	ViT-Base†	89.33	86.77	80.17	98.84	42.79	47.10	93.86
	ViT-Large	**90.07**	**87.57**	**80.90**	**99.02**	**44.29**	**48.36**	**94.37**

Table A4. Ablations of Dinomaly elements on VisA (%). NB: Noisy Bottleneck. LA: Linear Attention. LC: Loosen Constraint (2 groups). LL: Loosen Loss.

NB	LA	LC	LL	Image-level			Pixel-level			
				AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
				95.81	96.35	92.06	97.97	47.88	52.55	93.43
✓				97.38	97.74	94.07	97.84	50.42	54.57	93.71
	✓			95.74	96.23	91.87	98.01	47.89	52.58	93.34
		✓		96.39	97.01	92.54	97.37	46.80	51.66	92.75
			✓	96.93	97.26	93.32	98.37	49.52	53.59	94.11
✓	✓			97.52	97.75	94.33	98.06	51.49	55.09	93.75
✓		✓		98.06	98.37	95.18	98.21	51.43	54.89	93.94
✓		✓	✓	98.57	98.77	95.75	98.57	52.29	55.38	94.28
✓	✓	✓		98.22	98.43	95.27	98.51	53.11	55.48	94.24
✓	✓	✓	✓	**98.73**	**98.87**	**96.18**	**98.74**	**53.23**	**55.69**	**94.50**

of DINOv2 that employs 4 extra register tokens.

It is noted that most models are pre-trained with the image resolution of 224×224, except that DINOv2 [39] and DINOv2-R [7] have extra a high-resolution training phase with 518×518. Directly using the pre-trained weights on a different resolution for UAD without fine-tuning 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., the input size is 392×392 for ViT-Base/14 and 448×448 for ViT-Base/16. Additionally, we train Dinomaly with the low-resolution input size of 224×224.

The results are presented in Table A1. Within Dinomaly, nearly all foundation models can produce SoTA-level results with image-level AUROC higher than 98%. Generally speaking, CL+MIM combined models outperform MIM and CL models. In addition, most foundations do not benefit from a higher resolution for image-level performance but suffer from it, indicating the lack of generalization on a input size different from pre-training; while as expected, DINOv2 and DINOv2-R pre-trained on larger inputs can better benefit from higher resolution in Dinomaly. Because some methods, i.e., D-iGPT, DINO, and iBOT, produce similar results to DINOv2 in 224×224, we expect that they also have the potential to be as powerful in Dinomaly if they are pre-trained in high-resolution. Employing MAE produces the worst results. MAE was also tested as the backbone of ViTAD[65], resulting in undesirable performances (I-AUROC=95.3), which was attributed to the weak semantic expression caused by the pretraining strategy. It is also noted that MAE is bad in other unsupervised tasks such as ImageNet kNN; therefore, MAE is considered to be less effective in tasks without finetuning.

Input Size. The patch size of ViTs (usually 14×14 or 16×16) is much larger than the stem layer's down-sampling rate of CNNs (usually 4×4), resulting in smaller feature map size. For dense prediction tasks like semantic segmentation, ViTs usually employ a large input image size [39]. This practice holds in anomaly localization as well. In Table A2, we present the results of Dinomaly with different input resolutions. Following PatchCore [45], by default, we adopt center-crop preprocessing to reduce the influence of background, which can also cause unreachable anomalies at the edge of images. Experimental results demonstrate our robustness to input size. While small image size is enough for image-level anomaly detection, larger inputs are beneficial to anomaly localization. All experiments evaluate localization performance in a unified size of 256×256 for fairness.

Scalability on VisA and Real-IAD. We demonstrate the performance of different ViT sizes on VisA and Real-IAD in Table A3.

表A3. ViT架构在VisA和Real-IAD上的扩展性（%）。†：默认设置。

Dateset	Arch.	Image-level			Pixel-level			
		AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
VisA [70]	ViT-Small	97.94	98.09	95.33	98.57	51.19	55.10	93.71
	ViT-Base†	98.73	98.87	**96.18**	98.74	53.23	55.69	94.50
	ViT-Large	**98.85**	**99.09**	96.12	**99.10**	**55.68**	**57.33**	**94.76**
Real-IAD [54]	ViT-Small	89.10	86.91	79.87	98.69	41.88	46.74	94.08
	ViT-Base†	89.33	86.77	80.17	98.84	42.79	47.10	93.86
	ViT-Large	**90.07**	**87.57**	**80.90**	**99.02**	**44.29**	**48.36**	**94.37**

表A4. VisA上Dinomaly各要素的消融实验（%）。NB：噪声瓶颈。LA：线性注意力。LC：宽松约束（2组 ）。LL：松弛损失。

NB	LA	LC	LL	Image-level			Pixel-level			
				AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
				95.81	96.35	92.06	97.97	47.88	52.55	93.43
✓				97.38	97.74	94.07	97.84	50.42	54.57	93.71
	✓			95.74	96.23	91.87	98.01	47.89	52.58	93.34
		✓		96.39	97.01	92.54	97.37	46.80	51.66	92.75
			✓	96.93	97.26	93.32	98.37	49.52	53.59	94.11
✓	✓			97.52	97.75	94.33	98.06	51.49	55.09	93.75
✓		✓		98.06	98.37	95.18	98.21	51.43	54.89	93.94
✓		✓	✓	98.57	98.77	95.75	98.57	52.29	55.38	94.28
✓	✓	✓		98.22	98.43	95.27	98.51	53.11	55.48	94.24
✓	✓	✓	✓	**98.73**	**98.87**	**96.18**	**98.74**	**53.23**	**55.69**	**94.50**

DINOv2 采用了 4 个额外寄存器令牌的版本。

值得注意的是，大多数模型是以224×224的图像分辨率进行预训练的，但DINOv2 [39]和DINOv2-R [7]除外，它们额外进行了518×518的高分辨率训练阶段。与其他监督任务不同，在无监督异常检测中直接使用不同分辨率下的预训练权重而不进行微调，可能导致泛化问题。因此，默认情况下，我们仍将所有对比模型的特征尺寸保持在28×28，即ViT-Base/14的输入尺寸为392×392，ViT-Base/16的输入尺寸为448×448。此外，我们使用224×224的低分辨率输入尺寸训练Dinomaly。

结果呈现在表A1中。在Dino-maly框架内，几乎所有基础模型都能达到SoTA级别的性能，图像级AUROC超过98%。总体而言，CL+MIM组合模型的表现优于单独的MIM和CL模型。此外，大多数基础模型在图像级性能上并未从更高分辨率中受益，反而因此受损，这表明其在不同于预训练尺寸的输入上缺乏泛化能力；而正如预期，基于更大输入预训练的DINOv2和DINOv2-R在Dino-maly中能更好地利用高分辨率优势。由于某些方法（如D-iGPT、DINO和iBOT）在224×224分辨率下取得了与DINOv2相似的结果，我们预期若它们能在高分辨率下进行预训练，同样具备在Dinomaly中达到同等强大性能的潜力。

使用MAE产生了最差的结果。MAE也曾作为ViTAD[65]的骨干网络进行测试，但表现不佳（I-AUROC=95.3），这归因于其预训练策略导致的语义表达能力较弱。同时值得注意的是，MAE在其他无监督任务（如ImageNet kNN）中同样表现不佳；因此，MAE被认为在未经微调的任务中效果较差。

输入尺寸。ViT的补丁尺寸（通常为14×14或16×16）远大于CNN主干层的下采样率（通常为4×4），导致特征图尺寸更小。对于语义分割等密集预测任务，ViT通常采用较大的输入图像尺寸[39]。这一做法在异常定位任务中同样适用。在表A2中，我们展示了Dinomaly在不同输入分辨率下的结果。遵循PatchCore[45]的默认设置，我们采用中心裁剪预处理以减少背景干扰，这也能避免图像边缘的异常无法被检测的问题。实验结果表明我们的方法对输入尺寸具有鲁棒性。虽然小尺寸图像足以完成图像级异常检测，但更大的输入尺寸对异常定位更为有利。为公平起见，所有实验均在256×256的统一尺寸下评估定位性能。

在VisA和Real-IAD上的可扩展性。我们在表A3中展示了不同ViT尺寸在VisA和Real-IAD上的性能表现。

Table A5. Ablations of Dropout rates in Noisy Bottleneck, conducted on MVTec-AD (%). †: default.

Dropout rate	Image-level			Pixel-level			
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
0 (noiseless)	99.19	99.55	98.51	97.55	63.11	64.39	93.33
0.1	99.54	99.75	98.90	**98.35**	**69.46**	**69.19**	94.53
0.2 †	99.60	99.78	99.04	**98.35**	69.29	69.17	**94.79**
0.3	**99.65**	**99.83**	99.16	98.34	68.46	68.81	94.63
0.4	99.64	99.80	**99.23**	98.22	67.95	68.33	94.57
0.5	99.56	99.81	99.14	98.15	67.43	67.82	94.64

Table A6. Ablations of reconstruction constraint, conduected on MVTec-AD (%). †: default.

Constraints	Image-level			Pixel-level			
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
layer-to-layer (dense, every 1)	99.39	99.68	98.73	98.12	68.55	68.63	94.28
layer-to-layer (sparse, every 2)	99.52	99.73	98.95	98.16	68.89	68.57	94.40
layer-to-layer (sparse, every 4)	99.54	99.77	99.05	98.04	66.69	67.17	94.07
layer-to-cat-layer (every 2)	99.48	99.71	99.26	97.83	62.29	62.91	93.16
group-to-group (1 group)	**99.64**	**99.80**	**99.36**	98.18	64.79	65.40	93.96
group-to-group (2 groups)†	99.60	99.78	99.04	**98.35**	**69.29**	**69.17**	**94.79**

Table A7. Comparison between Convolutional block, Softmax Attention, and Linear Attention as the spatial mixer of decoder, conducted on MVTec-AD (%).

Spatial Mixer	Image-level			Pixel-level			
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
ConvBlock 3×3	99.45	99.63	98.64	98.05	65.35	68.07	94.17
ConvBlock 5×5	99.41	99.62	98.86	97.99	66.64	67.47	94.24
ConvBlock 7×7	99.42	99.65	98.86	98.01	67.57	67.94	94.45
Softmax Attention	99.52	99.73	98.92	98.20	68.25	68.34	94.17
Softmax Attention w/ Neighbour-Mask $n = 1$	99.51	99.71	98.90	98.17	67.86	67.92	94.27
Softmax Attention w/ Neighbour-Mask $n = 3$	99.56	99.76	99.05	98.28	69.26	68.17	94.50
Linear Attention	**99.60**	99.78	99.04	98.35	69.29	69.17	**94.79**
Linear Attention w/ Neighbour-Mask $n = 1$	**99.60**	99.78	99.04	98.32	68.77	68.72	94.75
Linear Attention w/ Neighbour-Mask $n = 3$	**99.60**	**99.80**	**99.14**	**98.38**	**69.65**	**69.38**	94.70

Ablations on VisA. Similar to Table 3 that conduct ablation experiments on MVTec-AD, we additionally run them on VisA for further validations. As shown in Table A4, proposed components of Dinomaly contribute to the AD performances on VisA as on MVTec-AD.

Noisy Rates. We conduct ablations on the discarding rate of the Dropouts in MLP bottleneck, as shown in Table A5. Experimental results demonstrate that Dinomaly is robust to different levels of dropout rate.

Reconstruction Constraint. We quantitatively examine different reconstruction schemes presented in Figure 4. As shown in Table A6, group-to-group LC outperforms layer-to-layer supervision. On image-level metrics, 1-group LC with all layers added performs similarly to its 2-group counterpart that separates low-level and high-level layers; however, 1-group LC mixes low-level and high-level features which is harmful for anomaly localization. More ablations on scalability, input size, pre-trained foundations, etc., are presented in Appendix C.

Attention *vs.* Convolution. Previous works and this paper have proposed to leverage attentions instead of convolutions in UAD. Here, we conduct experiments substituting the attention in the decoder of Dinomaly by convolutions as the spatial mixers. Following MetaFormer [61], we employ Inverted Bottleneck block that consists of 1×1 conv, GELU activation, $N \times N$ deep-wise conv, and 1×1 conv, sequentially. The results are shown in Table A7, where Attentions outperform Convolutions, especially for pixel-level

表A5. 在MVTec-AD上进行的噪声瓶颈中Dropout率消融实验（%）。†：默认设置。

Dropout rate	Image-level			Pixel-level			
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
0 (noiseless)	99.19	99.55	98.51	97.55	63.11	64.39	93.33
0.1	99.54	99.75	98.90	**98.35**	**69.46**	**69.19**	94.53
0.2 †	99.60	99.78	99.04	**98.35**	69.29	69.17	**94.79**
0.3	**99.65**	**99.83**	99.16	98.34	68.46	68.81	94.63
0.4	99.64	99.80	**99.23**	98.22	67.95	68.33	94.57
0.5	99.56	99.81	99.14	98.15	67.43	67.82	94.64

表A6. 重建约束的消融实验，在MVTec-AD上进行（%）。†：默认设置。

Constraints	Image-level			Pixel-level			
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
layer-to-layer (dense, every 1)	99.39	99.68	98.73	98.12	68.55	68.63	94.28
layer-to-layer (sparse, every 2)	99.52	99.73	98.95	98.16	68.89	68.57	94.40
layer-to-layer (sparse, every 4)	99.54	99.77	99.05	98.04	66.69	67.17	94.07
layer-to-cat-layer (every 2)	99.48	99.71	99.26	97.83	62.29	62.91	93.16
group-to-group (1 group)	**99.64**	**99.80**	**99.36**	98.18	64.79	65.40	93.96
group-to-group (2 groups)†	99.60	99.78	99.04	**98.35**	**69.29**	**69.17**	**94.79**

表A7。解码器空间混合器中卷积块、Softmax注意力与线性注意力的比较，实验 d
在MVTec-AD数据集上（%）。

Spatial Mixer	Image-level			Pixel-level			
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
ConvBlock 3×3	99.45	99.63	98.64	98.05	65.35	68.07	94.17
ConvBlock 5×5	99.41	99.62	98.86	97.99	66.64	67.47	94.24
ConvBlock 7×7	99.42	99.65	98.86	98.01	67.57	67.94	94.45
Softmax Attention	99.52	99.73	98.92	98.20	68.25	68.34	94.17
Softmax Attention w/ Neighbour-Mask $n = 1$	99.51	99.71	98.90	98.17	67.86	67.92	94.27
Softmax Attention w/ Neighbour-Mask $n = 3$	99.56	99.76	99.05	98.28	69.26	68.17	94.50
Linear Attention	**99.60**	99.78	99.04	98.35	69.29	69.17	**94.79**
Linear Attention w/ Neighbour-Mask $n = 1$	**99.60**	99.78	99.04	98.32	68.77	68.72	94.75
Linear Attention w/ Neighbour-Mask $n = 3$	**99.60**	**99.80**	**99.14**	**98.38**	**69.65**	**69.38**	94.70

VisA上的消融实验。类似于表3中对MVTec-AD进行的消融实验，我们额外在VisA数据集上进行了验证。如表A4所示，Dinomaly提出的各项组件在VisA数据集上对异常检测性能的提升作用与在MVTec-AD上表现一致。

噪声率。我们对MLP瓶颈中Dropout的丢弃率进行了消融实验，如表A5所示。实验结果表明，Dinomaly对不同水平的丢弃率具有鲁棒性。

重建约束。我们定量检验了图4中展示的不同重建方案。如表A6所示，组到组LC优于层到层监督。在图像级指标上，添加所有层的1组LC与其2组对应方案表现相似——对应部分将低层与高层特征分离；然而，单组LC混合了低层与高层特征，这对异常定位是有害的。关于可扩展性、输入尺寸、预训练基础模型等的更多消融实验详见附录C。

注意力 ***vs.*** 卷积。先前的研究与本文均提出在无监督异常检测中利用注意力机制替代卷积操作。我们通过实验将Dinomaly解码器中的注意力模块替换为卷积作为空间混合器。参照MetaFormer [61]的设计，我们采用由1×1卷积、GELU激活函数、$N \times N$深度可分离卷积及1×1卷积依次堆叠而成的倒残差模块。表A7结果显示注意力机制优于卷积方法，尤其在像素级检测任务中表现更为突出。

Table A8. Dropout *vs.* feature jitter, conducted on MVTec-AD (%).

Noise type	Image-level			Pixel-level			
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
No Noise	99.19	99.55	98.51	97.55	63.11	64.39	93.33
Patch Masking p=0.1	99.27	99.60	98.80	97.92	67.15	66.90	94.18
Patch Masking p=0.2	99.17	99.56	98.59	97.75	66.55	66.32	94.11
Patch Masking p=0.3	99.11	99.54	98.37	97.53	65.48	65.96	93.84
Patch Masking p=0.4	99.20	99.59	98.53	97.71	65.58	66.36	94.15
Feature Jitter *scale*=1	99.23	99.54	98.48	97.58	63.22	64.31	93.55
Feature Jitter *scale*=5	99.24	99.57	98.55	97.84	65.28	65.81	93.75
Feature Jitter *scale*=10	99.46	99.73	99.12	98.19	67.59	67.80	94.19
Feature Jitter *scale*=20	99.59	99.79	99.04	98.23	67.93	68.21	94.40
Dropout p=0.1	99.54	99.75	98.90	**98.35**	**69.46**	**69.19**	94.53
Dropout p=0.2	99.60	99.78	99.04	**98.35**	69.29	69.17	**94.79**
Dropout p=0.3	**99.65**	**99.83**	99.16	98.34	68.46	68.81	94.63
Dropout p=0.4	99.64	99.80	**99.23**	98.22	67.95	68.33	94.57

Table A9. Integrating the essense of Noisy Bottleneck (NB) and Loose Loss (LL) on RD4AD, conducted on MVTec-AD (%). †: Reproduction in our framework; ReLU in ResNet decoder is replaced by GELU, StableAdamW optimizer is used.

Method	NB	LL	Image-level			Pixel-level			
			AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
RD4AD†			97.8	99.1	97.2	96.4	58.0	59.3	91.9
RD4AD	✓		98.4	99.4	97.9	97.2	58.6	60.4	92.9
RD4AD		✓	98.2	99.2	97.5	96.8	60.0	61.1	92.7
RD4AD	✓	✓	98.5	99.4	97.8	97.2	59.6	61.2	93.0

Table A10. Scaling properties of a previous ViT-based method, ViTAD[65] on MVTec-AD. †: their original setting.

Method	Pre-Train Backbone	Input Size	Image-level			Pixel-level			
			AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
ViTAD†	DINO	256^2	98.3	99.4	97.3	97.7	55.3	58.7	91.4
ViTAD	MAE	256^2	95.3	97.7	95.2	97.4	53.0	56.2	90.6
ViTAD	DINOv2	256^2	98.7	99.4	98.1	97.6	55.3	59.1	92.7
ViTAD	DINOv2-R	256^2	98.5	99.3	97.8	97.4	54.5	59.2	92.8
ViTAD†	DINO	256^2	98.3	99.4	97.3	97.7	55.3	58.7	91.4
ViTAD	DINO	320^2	98.3	99.2	97.1	97.6	61.3	63.3	92.4
ViTAD	DINO	384^2	97.8	98.9	96.3	97.5	62.5	63.7	92.4

anomaly localization. In addition, utilizing convolutions in the decoder can still yield SoTA results, demonstrating the universality of the proposed Dinomaly.

Neighbour-Masking. Prior method [60] proposed to mask the keys and values in an $n \times n$ square centered at each query, in order to alleviate identity mapping in Attention. This mechanism can also be applied to Linear Attention as well. As shown in Table A7, neighbor-masking can further improve Dinomaly with both Softmax Attention and Linear Attention moderately.

Noise Bottleneck. UniAD [60] proposed to perturb the encoder features by Feature Jitter, i.e. adding Gaussian noise with *scale* to control the noise magnitude. It is also easy to borrow the masking strategy of MAE [19] to randomly mask patch tokens before the decoder. We evaluate the effectiveness of feature jitter and patch-masking in Dinomaly by placing it at the beginning of Noisy Bottleneck. As shown in Table A8, both Dropout and Feature Jitter can

表 A8. 在 MVTec-AD (%) 上进行的 Dropout *vs.* 特征抖动。

Noise type	Image-level			Pixel-level			
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
No Noise	99.19	99.55	98.51	97.55	63.11	64.39	93.33
Patch Masking p=0.1	99.27	99.60	98.80	97.92	67.15	66.90	94.18
Patch Masking p=0.2	99.17	99.56	98.59	97.75	66.55	66.32	94.11
Patch Masking p=0.3	99.11	99.54	98.37	97.53	65.48	65.96	93.84
Patch Masking p=0.4	99.20	99.59	98.53	97.71	65.58	66.36	94.15
Feature Jitter *scale*=1	99.23	99.54	98.48	97.58	63.22	64.31	93.55
Feature Jitter *scale*=5	99.24	99.57	98.55	97.84	65.28	65.81	93.75
Feature Jitter *scale*=10	99.46	99.73	99.12	98.19	67.59	67.80	94.19
Feature Jitter *scale*=20	99.59	99.79	99.04	98.23	67.93	68.21	94.40
Dropout p=0.1	99.54	99.75	98.90	**98.35**	**69.46**	**69.19**	94.53
Dropout p=0.2	99.60	99.78	99.04	**98.35**	69.29	69.17	**94.79**
Dropout p=0.3	**99.65**	**99.83**	99.16	98.34	68.46	68.81	94.63
Dropout p=0.4	99.64	99.80	**99.23**	98.22	67.95	68.33	94.57

表A9. 在RD4AD上整合噪声瓶颈（NB）与宽松损失（LL）的核心思想，在MVTec-AD数据集上的实验结果（%）。†：代表哦-在我们的框架中进行了调整；ResNet解码器中的ReLU被替换为GELU，并采用了StableAdamW优化器。

Method	NB	LL	Image-level			Pixel-level			
			AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
RD4AD†			97.8	99.1	97.2	96.4	58.0	59.3	91.9
RD4AD	✓		98.4	99.4	97.9	97.2	58.6	60.4	92.9
RD4AD		✓	98.2	99.2	97.5	96.8	60.0	61.1	92.7
RD4AD	✓	✓	98.5	99.4	97.8	97.2	59.6	61.2	93.0

表 A10. 缩放属性　　先前基于ViT的方法ViTAD[65]在MVTec-AD上的性能。†：第　其原始设定。

Method	Pre-Train Backbone	Input Size	Image-level			Pixel-level			
			AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
ViTAD†	DINO	256^2	98.3	99.4	97.3	97.7	55.3	58.7	91.4
ViTAD	MAE	256^2	95.3	97.7	95.2	97.4	53.0	56.2	90.6
ViTAD	DINOv2	256^2	98.7	99.4	98.1	97.6	55.3	59.1	92.7
ViTAD	DINOv2-R	256^2	98.5	99.3	97.8	97.4	54.5	59.2	92.8
ViTAD†	DINO	256^2	98.3	99.4	97.3	97.7	55.3	58.7	91.4
ViTAD	DINO	320^2	98.3	99.2	97.1	97.6	61.3	63.3	92.4
ViTAD	DINO	384^2	97.8	98.9	96.3	97.5	62.5	63.7	92.4

异常定位。此外，在解码器中利用卷积仍能取得最先进的结果，这证明了所提出的Dinomaly的普适性。

邻域掩码。先前的方法[60]提出以每个查询为中心，对{v*}方形区域内的键和值进行掩码，以缓解注意力机制中的恒等映射问题。该机制同样适用于线性注意力。如表A7所示，邻域掩码能进一步提升Dinomaly在Softmax注意力和线性注意力适度。

噪声瓶颈。UniAD [60] 提出通过特征抖动（即添加高斯噪声，其幅度由 {v*} 控制）来扰动编码器特征。借鉴 MAE [19] 的掩码策略，在解码器前随机掩码图像块标记也较为简便。我们通过在噪声瓶颈起始处引入特征抖动与图像块掩码，评估了它们在 Di-nomaly 中的有效性。如表 A8 所示，Dropout 与特征抖动均能

Table A11. Matching previous methods in computation consumption. Dinomaly can be easily scaled by model size and input size.

Method	Params	MACs	MVTec-AD [3]			VisA [70]		
			I-AUROC	P-AUROC	P-AUPRO	I-AUROC	P-AUROC	P-AUPRO
DiAD [18]	1331M	451.5G	97.2	96.8	90.7	86.8	96.0	75.2
ReContrast [14]	154.2M	67.4G	98.3	97.1	93.2	95.5	98.5	91.9
RD4AD [10]	126.7M	32.1G	94.6	96.1	91.1	92.4	98.1	91.8
ViTAD [65]	39.0M	9.7G	98.3	97.7	91.4	90.5	98.2	85.1
Dinomaly-Base-392^2	148M	104.7G	99.6	98.4	94.8	98.7	98.7	94.5
Dinomaly-Base-280^2	148M	53.7G	99.6	98.2	93.6	97.8	98.7	92.4
Dinomaly-Small-392^2	37.4M	26.2G	99.3	98.1	94.4	97.9	98.6	93.7
Dinomaly-Small-280^2	37.4M	14.5G	99.3	98.0	93.4	96.5	98.5	90.9

Table A12. Results of 5 random seeds on MVTec-AD (%).

Random Seed	Image-level			Pixel-level			
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
seed=1	99.60	99.78	99.04	98.35	69.29	69.17	94.79
seed=2	99.63	99.79	99.12	98.33	68.73	68.91	94.63
seed=3	99.63	99.79	99.16	98.31	68.70	68.93	94.60
seed=4	99.56	99.74	99.02	98.33	69.04	69.09	94.70
seed=5	99.59	99.77	99.02	98.32	68.64	68.47	94.51
mean±std	99.60±0.03	99.77±0.02	99.07±0.06	98.33±0.01	68.88±0.25	68.91±0.24	94.65±0.09

be a good noise injector in Noisy Bottleneck. Meanwhile, Dropout is more robust to the noisy scale hyperparameter, and more elegant without introducing new modules.

Adaptation on CNN Method. Some proposed elements (Linear Attention and Loose Constraint) are closely bounded to modern ViTs. Loose Loss (hard-mining) can be directly applied to previous CNN-based methods, e.g., RD4AD [10]. Noisy Bottleneck can be adapted to RD4AD with minor modifications (apply dropout before MFF layer). We apply these modules to RD4AD to validate the effectiveness of our contributions. The results are shown in Table A9, where these two elements boost the performance of RD4AD to a whole new level that can be compared with prior MUAD SoTAs.

Scaling of Compared Method. As previously discussed in the Experiment section, compared method cannot fully utilize the scaling of pre-trained method, model size, and input size. For example, RD4AD [10] found WideResNet50 better than WideResNet101 as the encoder backbone. ViTAD [65] found ViT-Small better than ViT-Base. Here, we also present the experiments on pre-training method and input size of ViTAD, as shown in Table A10. It is also noted that the paradigm of ViTAD is very similar to RD4AD (replacing CNN by ViT) as well as the starting point of Dinomaly (the first row in the ablation Table 3).

Computation Comparison. The computation costs of Dinomaly variants were previously presented in Table 4 and Table A2. Here, we compare the computation consumption of Dinomaly and prior works. As shown in Table A11, Dinomaly can be easily scaled by model size and input size to match different application scenarios.

Random Seeds. Due to limited computation resources, experiments in this paper are conducted for one run with random seed=1. Here, we conduct 5 runs with 5 random seeds on MVTec-AD. As shown in Table A12, Dinomaly is robust to randomness.

D. Additional Dataset

To demonstrate the generalization of our method, we conduct experiments on three more popular anomaly detection datasets under MUAD setting, including MPDD and BTAD and Uni-Medical. The MPDD [24] (Metal Parts Defect Detection Dataset) is a dataset aimed at benchmarking visual defect detection methods in industrial metal parts manufacturing. It consists of more than 1346 images across 6 categories with pixel-precise defect annotation masks. The BTAD [38] (beanTech Anomaly Detection) dataset is a real-world industrial anomaly dataset. The dataset contains a total of 2830 real-world images of 3 industrial products showcasing body and surface defects. It is noted that the training set of BTAD is noisy because it contains anomalous samples [25]. Uni-Medical [66] is a medical UAD dataset consisting of 2D image slices from 3D CT volumes. It con-

表A11. 在计算消耗方面与先前方法的匹配情况。Dinomaly可通过模型规模和输入规模轻松扩展。

Method	Params	MACs	MVTec-AD [3]			VisA [70]		
			I-AUROC	P-AUROC	P-AUPRO	I-AUROC	P-AUROC	P-AUPRO
DiAD [18]	1331M	451.5G	97.2	96.8	90.7	86.8	96.0	75.2
ReContrast [14]	154.2M	67.4G	98.3	97.1	93.2	95.5	98.5	91.9
RD4AD [10]	126.7M	32.1G	94.6	96.1	91.1	92.4	98.1	91.8
ViTAD [65]	39.0M	9.7G	98.3	97.7	91.4	90.5	98.2	85.1
Dinomaly-Base-392^2	148M	104.7G	99.6	98.4	94.8	98.7	98.7	94.5
Dinomaly-Base-280^2	148M	53.7G	99.6	98.2	93.6	97.8	98.7	92.4
Dinomaly-Small-392^2	37.4M	26.2G	99.3	98.1	94.4	97.9	98.6	93.7
Dinomaly-Small-280^2	37.4M	14.5G	99.3	98.0	93.4	96.5	98.5	90.9

表A12. 在MVTec-AD数据集上5个随机种子的结果（%）。

Random Seed	Image-level			Pixel-level			
	AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
seed=1	99.60	99.78	99.04	98.35	69.29	69.17	94.79
seed=2	99.63	99.79	99.12	98.33	68.73	68.91	94.63
seed=3	99.63	99.79	99.16	98.31	68.70	68.93	94.60
seed=4	99.56	99.74	99.02	98.33	69.04	69.09	94.70
seed=5	99.59	99.77	99.02	98.32	68.64	68.47	94.51
mean±std	99.60±0.03	99.77±0.02	99.07±0.06	98.33±0.01	68.88±0.25	68.91±0.24	94.65±0.09

在噪声瓶颈中成为一个优秀的噪声注入器。同时，Dropout对噪声尺度超参数具有更强的鲁棒性，且无需引入新模块，设计更为优雅。

CNN方法的适配。一些提出的元素（线性注意力与宽松约束）与现代ViTs紧密相关。宽松损失（困难样本挖掘）可直接应用于先前的基于CNN的方法，例如RD4AD [10]。噪声瓶颈可通过微小修改（在MFF层前应用dropout）适配到RD4AD。我们将这些模块应用于RD4AD以验证我们贡献的有效性。结果如表A9所示，这两个元素将RD4AD的性能提升至全新水平，可与先前MUAD的先进方法相媲美。

对比方法的缩放性。如实验部分先前讨论，对比方法无法充分利用预训练方法、模型尺寸和输入尺寸的缩放优势。例如，RD4AD [10] 发现WideResNet50作为编码器骨干网络优于WideResNet101。ViTAD [65] 发现ViT-Small优于ViT-Base。此处，我们也展示了ViTAD在预训练方法和输入尺寸上的实验，如表A10所示。同时值得注意的是，ViTAD的范式与RD4AD（用ViT替换CNN）以及Dino-maly的起点（消融实验表3中的第一行）非常相似。

计算比较。Dinomaly变体的计算成本先前已在表4中呈现，且表A2。此处，我们对比了Dinomaly与先前工作的计算消耗。如表A11所示，Dinomaly可通过模型规模和输入尺寸轻松扩展，以适应不同的应用场景。

随机种子。由于计算资源有限，本文中的实验仅使用随机种子=1进行了一次运行。此处，我们在MVTec-AD数据集上使用5个随机种子进行了5次运行。如表A12所示，Dinomaly对随机性具有鲁棒性。

D. 附加数据集

为了证明我们方法的泛化能力，我们在MUAD设置下对另外三个流行的异常检测数据集进行了实验，包括MPDD、BTAD和Uni-Medical。MPDD[24]（金属零件缺陷检测数据集）是一个旨在为工业金属零件制造中的视觉缺陷检测方法提供基准的数据集。它包含超过1346张图像，涵盖6个类别，并带有像素级精确的缺陷标注掩码。BTAD[38]（beanTech异常检测）数据集是一个真实世界的工业异常数据集。该数据集总共包含2830张真实世界图像，展示了3种工业产品的本体和表面缺陷。值得注意的是，BTAD的训练集存在噪声，因为它包含异常样本[25]。Uni-Medical[66]是一个医学UAD数据集，由来自3D CT体积的2D图像切片组成。它包

Table A13. Performance on MPDD and BTAD under **multi-class** UAD setting (%). †: method designed for MUAD.

Dateset	Method	Image-level			Pixel-level			
		AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
MPDD [24]	RD4AD [10]	90.3	92.8	90.5	98.3	39.6	40.6	95.2
	SimpleNet [34]	90.6	94.1	89.7	97.1	33.6	35.7	90.0
	DeSTSeg [67]	92.6	91.8	92.8	90.8	30.6	32.9	78.3
	UniAD [60]†	80.1	83.2	85.1	95.4	19.0	25.6	83.8
	DiAD [18]†	85.8	89.2	86.5	91.4	15.3	19.2	66.1
	ViTAD [65]†	87.4	90.8	87.0	97.8	44.1	46.4	95.3
	MambaAD [17]†	89.2	93.1	90.3	97.7	33.5	38.6	92.8
	Dinomaly (Ours)	**97.2**	**98.4**	**96.0**	**99.1**	**59.5**	**59.4**	**96.6**
BTAD [38]	RD4AD [10]	94.1	96.8	93.8	**98.0**	57.1	58.0	**79.9**
	SimpleNet [34]	94.0	97.9	93.9	96.2	41.0	43.7	69.6
	DeSTSeg [67]	93.5	96.7	93.8	94.8	39.1	38.5	72.9
	UniAD [60]†	94.5	**98.4**	94.9	97.4	52.4	55.5	78.9
	DiAD [18]†	90.2	88.3	92.6	91.9	20.5	27.0	70.3
	ViTAD [65]†	94.0	97.0	93.7	97.6	58.3	56.5	72.8
	MambaAD [17]†	92.9	96.2	93.0	97.6	51.2	55.1	77.3
	Dinomaly (Ours)	**95.4**	**98.4**	**95.6**	97.8	**70.1**	**68.0**	76.5
Uni-Medical [66]	RD4AD [10]	76.1	75.3	78.2	96.5	38.3	39.8	86.8
	SimpleNet [34]	77.5	77.7	76.7	94.3	34.4	36.0	77.0
	DeSTSeg [67]	78.5	77.0	78.2	65.7	41.7	34.0	35.3
	UniAD [60]†	79.0	76.1	77.1	96.6	39.3	41.1	86.0
	DiAD [18]†	78.8	77.2	77.7	95.8	34.2	35.5	84.3
	ViTAD [65]†	81.8	80.7	80.0	**97.1**	48.3	48.2	86.7
	MambaAD [17]†	83.9	80.8	**81.9**	96.8	45.8	47.5	**88.2**
	Dinomaly (Ours)	**84.9**	**84.1**	81.0	96.8	**51.7**	**52.1**	85.5

tains 13339 training images and 7013 test images across three objects, i.e., brain CT, liver CT, and retinal OCT. This dataset is not entirely suitable for evaluating 2D anomaly detection methods, as identifying lesions in medical images requires 3D contextual information. The training hyperparameters are the same to MVTec-AD, except the dropout rate for Uni-Medical is increased to 0.4. The performance of Dinomaly and previous SoTAs is presented in Table A13, where our method demonstrates superior results.

E. Results Per-Category

For future research, we report the per-class results of MVTec-AD [3], VisA [70], and Real-IAD [54]. The performance of compared methods is drawn from MambaAD [17]. Thanks for their exhaustive reproducing. The results of image-level anomaly detection and pixel-level anomaly localization on MVTec-AD are presented in Table A14 and Table A15, respectively. The results of image-level anomaly detection and pixel-level anomaly localization on VisA are presented in Table A16 and Table A17, respectively. The results of image-level anomaly detection and pixel-level anomaly localization on Real-IAD are presented in Table A18 and Table A19, respectively.

F. Qualitative Visualization

We visualize the output anomaly maps of Dinomaly on MVTec-AD, VisA, and Real-IAD, as shown in Figure A1, Figure A2, and Figure A3. It is noted that all visualized samples are randomly chosen without artificial selection.

G. Limitation

Vision Transformers are known for their high computation cost, which can be a barrier to low-computation scenarios that require inference speed. Future research can be conducted on the efficiency of Transformer-based methods, such as distillation, pruning, and hardware-friendly attention mechanism (such as FlashAttention).

As discussed in section A, Dinomaly is used for sensory AD that aims to detect regional anomalies in normal backgrounds. It is not suitable for semantic AD. Previous works have shown that methods designed for sensory AD usually fail to be competitive under semantic AD tasks [10, 60]. Conversely, methods designed for semantic AD do not perform well on sensory AD tasks [42, 43]. Future work can be conducted to unify these two tasks, but according to the "no free lunch" theorem, we believe that methods designed for specific anomaly assumption are likely to be more convincing.

Other special UAD settings, such as zero-shot UAD (vision-language model based) [23], few-shot UAD [22], UAD under noisy training set [25], are not included in this work.

表A13. 多类别异常检测设置下在MPDD与BTAD数据集上的性能表现（%）。†：专为多类别异常检测设计的方法。

Dateset	Method	Image-level			Pixel-level			
		AUROC	AP	F_1-max	AUROC	AP	F_1-max	AUPRO
MPDD [24]	RD4AD [10]	90.3	92.8	90.5	98.3	39.6	40.6	95.2
	SimpleNet [34]	90.6	94.1	89.7	97.1	33.6	35.7	90.0
	DeSTSeg [67]	92.6	91.8	92.8	90.8	30.6	32.9	78.3
	UniAD [60]†	80.1	83.2	85.1	95.4	19.0	25.6	83.8
	DiAD [18]†	85.8	89.2	86.5	91.4	15.3	19.2	66.1
	ViTAD [65]†	87.4	90.8	87.0	97.8	44.1	46.4	95.3
	MambaAD [17]†	89.2	93.1	90.3	97.7	33.5	38.6	92.8
	Dinomaly (Ours)	**97.2**	**98.4**	**96.0**	**99.1**	**59.5**	**59.4**	**96.6**
BTAD [38]	RD4AD [10]	94.1	96.8	93.8	**98.0**	57.1	58.0	**79.9**
	SimpleNet [34]	94.0	97.9	93.9	96.2	41.0	43.7	69.6
	DeSTSeg [67]	93.5	96.7	93.8	94.8	39.1	38.5	72.9
	UniAD [60]†	94.5	**98.4**	94.9	97.4	52.4	55.5	78.9
	DiAD [18]†	90.2	88.3	92.6	91.9	20.5	27.0	70.3
	ViTAD [65]†	94.0	97.0	93.7	97.6	58.3	56.5	72.8
	MambaAD [17]†	92.9	96.2	93.0	97.6	51.2	55.1	77.3
	Dinomaly (Ours)	**95.4**	**98.4**	**95.6**	97.8	**70.1**	**68.0**	76.5
Uni-Medical [66]	RD4AD [10]	76.1	75.3	78.2	96.5	38.3	39.8	86.8
	SimpleNet [34]	77.5	77.7	76.7	94.3	34.4	36.0	77.0
	DeSTSeg [67]	78.5	77.0	78.2	65.7	41.7	34.0	35.3
	UniAD [60]†	79.0	76.1	77.1	96.6	39.3	41.1	86.0
	DiAD [18]†	78.8	77.2	77.7	95.8	34.2	35.5	84.3
	ViTAD [65]†	81.8	80.7	80.0	**97.1**	48.3	48.2	86.7
	MambaAD [17]†	83.9	80.8	**81.9**	96.8	45.8	47.5	**88.2**
	Dinomaly (Ours)	**84.9**	**84.1**	81.0	96.8	**51.7**	**52.1**	85.5

该数据集包含13339张训练图像和7013张测试图像，涵盖三个对象，即脑部CT、肝脏CT和视网膜OCT。此数据集并不完全适用于评估2D异常检测方法，因为识别医学图像中的病变需要3D上下文信息。训练超参数与MVTec-AD相同，但Uni-Medical的丢弃率提高至0.4。Dinomaly与先前最优方法的性能呈现在表A13中，我们的方法展示了优越的结果。

E. 各类别结果

对于未来的研究，我们报告了MVTec-AD [3]、VisA [70]和Real-IAD [54]的各类别结果。对比方法的性能数据取自MambaAD [17]，感谢他们详尽的复现工作。MVTec-AD上图像级异常检测和像素级异常定位的结果分别呈现在表A14和表A15中。VisA上图像级异常检测和像素级异常定位的结果分别呈现在表A16和表A17中。Real-IAD上图像级异常检测和像素级异常定位的结果分别呈现在表A18和表A19中。

F. 定性可视化

我们在MVTec-AD、VisA和Real-IAD数据集上对Dinomaly的输出异常图进行了可视化，如图A1所示，图A2和图A3。值得注意的是，所有可视化样本均为随机选取，未经人工筛选。

G. 局限性

视觉变换器以其高计算成本而闻名，这可能会成为需要推理速度的低计算场景的障碍。未来的研究可以集中在基于变换器方法的效率上，例如蒸馏、剪枝和硬件友好的注意力机制（如FlashAttention）。

如A节所述，Dinomaly用于旨在检测正常背景中区域异常的感官异常检测。它不适用于语义异常检测。先前的研究表明，为感官异常检测设计的方法通常在语义异常检测任务中缺乏竞争力[10, 60]。反之，为语义异常检测设计的方法在感官异常检测任务上表现不佳[42, 43]。未来的工作可以尝试统一这两类任务，但根据“没有免费午餐”定理，我们认为针对特定异常假设设计的方法可能更具说服力。

其他特殊的无监督异常检测设置，如零样本无监督异常检测（基于视觉语言模型）[23]、少样本无监督异常检测[22]、噪声训练集下的无监督异常检测[25]等，未包含在本工作中。

Table A14. Per-class performance on **MVTec-AD** dataset for multi-class anomaly detection with AUROC/AP/F_1-max metrics.

	Method →	RD4AD [10]	UniAD [60]	SimpleNet [34]	DeSTSeg [67]	DiAD [18]	MambaAD [17]	Dinomaly
	Category ↓	CVPR'22	NeurIPS'22	CVPR'23	CVPR'23	AAAI'24	Arxiv'24	Ours
Objects	Bottle	99.6/99.9/98.4	99.7/**100./100.**	**100./100./100.**	98.7/99.6/96.8	99.7/96.5/91.8	**100./100./100.**	**100./100./100.**
	Cable	84.1/89.5/82.5	95.2/95.9/88.0	97.5/98.5/94.7	89.5/94.6/85.9	94.8/98.8/95.2	98.8/99.2/95.7	**100./100./100.**
	Capsule	94.1/96.9/96.9	86.9/97.8/94.4	90.7/97.9/93.5	82.8/95.9/92.6	89.0/97.5/95.5	94.4/98.7/94.9	**97.9/99.5/97.7**
	Hazelnut	60.8/69.8/86.4	99.8/**100.**/99.3	99.9/99.9/99.3	98.8/99.2/98.6	99.5/99.7/97.3	**100./100./100.**	**100./100./100.**
	Metal Nut	**100./100.**/99.5	99.2/99.9/99.5	96.9/99.3/96.1	92.9/98.4/92.2	99.1/96.0/91.6	99.9/**100.**/99.5	**100./100./100.**
	Pill	97.5/99.6/96.8	93.7/98.7/95.7	88.2/97.7/92.5	77.1/94.4/91.7	95.7/98.5/94.5	97.0/99.5/96.2	**99.1/99.9/98.3**
	Screw	97.7/99.3/95.8	87.5/96.5/89.0	76.7/90.6/87.7	69.9/88.4/85.4	90.7/**99.7/97.9**	94.7/97.9/94.0	**98.4**/99.5/96.1
	Toothbrush	97.2/99.0/94.7	94.2/97.4/95.2	89.7/95.7/92.3	71.7/89.3/84.5	99.7/99.9/99.2	98.3/99.3/98.4	**100./100./100.**
	Transistor	94.2/95.2/90.0	99.8/98.0/93.8	99.2/98.7/97.6	78.2/79.5/68.8	99.8/99.6/97.4	**100./100./100.**	99.0/98.0/96.4
	Zipper	99.5/99.9/99.2	95.8/99.5/97.1	99.0/99.7/98.3	88.4/96.3/93.1	95.1/99.1/94.4	99.3/99.8/97.5	**100./100./100.**
Textures	Carpet	98.5/99.6/97.2	**99.8**/99.9/**99.4**	95.7/98.7/93.2	95.9/98.8/94.9	99.4/99.9/98.3	**99.8**/99.9/**99.4**	**99.8/100.**/98.9
	Grid	98.0/99.4/96.5	98.2/99.5/97.3	97.6/99.2/96.4	97.9/99.2/96.6	98.5/99.8/97.7	**100./100./100.**	99.9/**100.**/99.1
	Leather	**100./100./100.**	**100./100./100.**	**100./100./100.**	99.2/99.8/98.9	99.8/99.7/97.6	**100./100./100.**	**100./100./100.**
	Tile	98.3/99.3/96.4	99.3/99.8/98.2	99.3/99.8/98.8	97.0/98.9/95.3	96.8/99.9/98.4	98.2/99.3/95.4	**100./100./100.**
	Wood	99.2/99.8/98.3	98.6/99.6/96.6	98.4/99.5/96.7	99.9/**100.**/99.2	99.7/**100./100.**	98.8/99.6/96.6	99.8/99.9/99.2
	Mean	94.6/96.5/95.2	96.5/98.8/96.2	95.3/98.4/95.8	89.2/95.5/91.6	97.2/99.0/96.5	98.6/99.6/97.8	**99.6/99.8/99.0**

Table A15. Per-class performance on **MVTec-AD** dataset for multi-class anomaly localization with AUROC/AP/F_1-max/AUPRO metrics.

	Method →	RD4AD [10]	UniAD [60]	SimpleNet [34]	DeSTSeg [67]	DiAD [18]	MambaAD [17]	Dinomaly
	Category ↓	CVPR'22	NeurIPS'22	CVPR'23	CVPR'23	AAAI'24	Arxiv'24	Ours
Objects	Bottle	97.8/68.2/67.6/94.0	98.1/66.0/69.2/93.1	97.2/53.8/62.4/89.0	93.3/61.7/56.0/67.5	98.4/52.2/54.8/86.6	98.8/79.7/76.7/95.2	**99.2/88.6/84.2/96.6**
	Cable	85.1/26.3/33.6/75.1	97.3/39.9/45.2/86.1	96.7/42.4/51.2/85.4	89.3/37.5/40.5/49.4	96.8/50.1/57.8/80.5	95.8/42.2/48.1/90.3	**98.6/72.0/74.3/94.2**
	Capsule	**98.8**/43.4/50.0/94.8	98.5/42.7/46.5/92.1	98.5/35.4/44.3/84.5	95.8/47.9/48.9/62.1	97.1/42.0/45.3/87.2	98.4/43.9/47.7/92.6	98.7/**61.4/60.3/97.2**
	Hazelnut	97.9/36.2/51.6/92.7	98.1/55.2/56.8/94.1	98.4/44.6/51.4/87.4	98.2/65.8/61.6/84.5	98.3/79.2/**80.4**/91.5	99.0/63.6/64.4/95.7	**99.4/82.2**/76.4/**97.0**
	Metal Nut	94.8/55.5/66.4/91.9	62.7/14.6/29.2/81.8	**98.0/83.1**/79.4/85.2	84.2/42.0/22.8/53.0	97.3/30.0/38.3/90.6	96.7/74.5/79.1/93.7	96.9/78.6/**86.7/94.9**
	Pill	97.5/63.4/65.2/95.8	95.0/44.0/53.9/95.3	96.5/72.4/67.7/81.9	96.2/61.7/41.8/27.9	95.7/46.0/51.4/89.0	97.4/64.0/66.5/95.7	**97.8/76.4/71.6/97.3**
	Screw	99.4/40.2/44.6/96.8	98.3/28.7/37.6/95.2	96.5/15.9/23.2/84.0	93.8/19.9/25.3/47.3	97.9/**60.6/59.6**/95.0	99.5/49.8/50.9/97.1	**99.6**/60.2/**59.6/98.3**
	Toothbrush	**99.0**/53.6/58.8/92.0	98.4/34.9/45.7/87.9	98.4/46.9/52.5/87.4	96.2/52.9/58.8/30.9	**99.0**/78.7/72.8/95.0	**99.0**/48.5/59.2/91.7	98.9/**51.5/62.6/95.3**
	Transistor	85.9/42.3/45.2/74.7	**97.9**/59.5/**64.6/93.5**	95.8/58.2/56.0/83.2	73.6/38.4/39.2/43.9	95.1/15.6/31.7/90.0	96.5/69.4/67.1/87.0	93.2/**59.9**/58.5/77.0
	Zipper	98.5/53.9/60.3/94.1	96.8/40.1/49.9/92.6	97.9/53.4/54.6/90.7	97.3/64.7/59.2/66.9	96.2/60.7/60.0/91.6	98.4/60.4/61.7/94.3	**99.2/79.5/75.4/97.2**
Textures	Carpet	99.0/58.5/60.4/95.1	98.5/49.9/51.1/94.4	97.4/38.7/43.2/90.6	93.6/59.9/58.9/89.3	98.6/42.2/46.4/90.6	99.2/60.0/63.3/96.7	**99.3/68.7/71.1/97.6**
	Grid	96.5/23.0/28.4/97.0	63.1/10.7/11.9/92.9	96.8/20.5/27.6/88.6/	97.0/42.1/46.9/86.8	96.6/66.0/64.1/94.0	99.2/47.4/47.7/97.0	**99.4/55.3/57.7/97.2**
	Leather	99.3/38.0/45.1/97.4	98.8/32.9/34.4/96.8	98.7/28.5/32.9/92.7	**99.5/71.5/66.5**/91.1	98.8/56.1/62.3/91.3	99.4/50.3/53.3/98.7	99.4/52.2/55.0/**97.6**
	Tile	95.3/48.5/60.5/85.8	91.8/42.1/50.6/78.4	95.7/60.5/59.9/90.6	93.0/71.0/66.2/87.1	92.4/65.7/64.1/**90.7**	93.8/45.1/54.8/80.0	**98.1/80.1/75.7**/90.5
	Wood	95.3/47.8/51.0/90.0	93.2/37.2/41.5/86.7	91.4/34.8/39.7/76.3	95.9/**77.3/71.3**/83.4	93.3/43.3/43.5/**97.5**	94.4/46.2/48.2/91.2	**97.6**/72.8/68.4/94.0
	Mean	96.1/48.6/53.8/91.1	96.8/43.4/49.5/90.7	96.9/45.9/49.7/86.5	93.1/54.3/50.9/64.8	96.8/52.6/55.5/90.7	97.7/56.3/59.2/93.1	**98.4/69.3/69.2/94.8**

Table A16. Per-class performance on **VisA** dataset for multi-class anomaly detection with AUROC/AP/F1-max metrics.

Method →	RD4AD [10]	UniAD [60]	SimpleNet [34]	DeSTSeg [67]	DiAD [18]	MambaAD	Dinomaly
Category ↓	CVPR'22	NeurIPS'22	CVPR'23	CVPR'23	AAAI'24	Arxiv'24	Ours
pcb1	96.2/95.5/91.9	92.8/92.7/87.8	91.6/91.9/86.0	87.6/83.1/83.7	88.1/88.7/80.7	95.4/93.0/91.6	**99.1/99.1/96.6**
pcb2	97.8/97.8/94.2	87.8/87.7/83.1	92.4/93.3/84.5	86.5/85.8/82.6	91.4/91.4/84.7	94.2/93.7/89.3	**99.3/99.2/97.0**
pcb3	96.4/96.2/91.0	78.6/78.6/76.1	89.1/91.1/82.6	93.7/95.1/87.0	86.2/87.6/77.6	93.7/94.1/86.7	**98.9/98.9/96.1**
pcb4	99.9/99.9/99.0	98.8/98.8/94.3	97.0/97.0/93.5	97.8/97.8/92.7	99.6/99.5/97.0	**99.9/99.9/98.5**	99.8/99.8/98.0
macaroni1	75.9/ 1.5/76.8	79.9/79.8/72.7	85.9/82.5/73.1	76.6/69.0/71.0	85.7/85.2/78.8	91.6/89.8/81.6	**98.0/97.6/94.2**
macaroni2	88.3/84.5/83.8	71.6/71.6/69.9	68.3/54.3/59.7	68.9/62.1/67.7	62.5/57.4/69.6	81.6/78.0/73.8	**95.9/95.7/90.7**
capsules	82.2/90.4/81.3	55.6/55.6/76.9	74.1/82.8/74.6	87.1/93.0/84.2	58.2/69.0/78.5	91.8/95.0/88.8	**98.6/99.0/97.1**
candle	92.3/92.9/86.0	94.1/94.0/86.1	84.1/73.3/76.6	94.9/94.8/89.2	92.8/92.0/87.6	96.8/96.9/90.1	**98.7/98.8/95.1**
cashew	92.0/95.8/90.7	92.8/92.8/91.4	88.0/91.3/84.7	92.0/96.1/88.1	91.5/95.7/89.7	94.5/97.3/91.1	**98.7/99.4/97.0**
chewinggum	94.9/97.5/92.1	96.3/96.2/95.2	96.4/98.2/93.8	95.8/98.3/94.7	99.1/99.5/95.9	97.7/98.9/94.2	**99.8/99.9/99.0**
fryum	95.3/97.9/91.5	83.0/83.0/85.0	88.4/93.0/83.3	92.1/96.1/89.5	89.8/95.0/87.2	95.2/97.7/90.5	**98.8/99.4/96.5**
pipe_fryum	97.9/98.9/96.5	94.7/94.7/93.9	90.8/95.5/88.6	94.1/97.1/91.9	96.2/98.1/93.7	98.7/99.3/97.0	**99.2/99.7/97.0**
Mean	92.4/92.4/89.6	85.5/85.5/84.4	87.2/87.0/81.8	88.9/89.0/85.2	86.8/88.3/85.1	94.3/94.5/89.4	**98.7/98.9/96.2**

表A14. 在MVTec-AD数据集上使用AUROC/AP/F_1-max指标进行多类别异常检测的各类别性能表现。

	Method →	RD4AD [10]	UniAD [60]	SimpleNet [34]	DeSTSeg [67]	DiAD [18]	MambaAD [17]	Dinomaly
	Category ↓	CVPR'22	NeurIPS'22	CVPR'23	CVPR'23	AAAI'24	Arxiv'24	Ours
Objects	Bottle	99.6/99.9/98.4	99.7/**100./100.**	**100./100./100.**	98.7/99.6/96.8	99.7/96.5/91.8	**100./100./100.**	**100./100./100.**
	Cable	84.1/89.5/82.5	95.2/95.9/88.0	97.5/98.5/94.7	89.5/94.6/85.9	94.8/98.8/95.2	98.8/99.2/95.7	**100./100./100.**
	Capsule	94.1/96.9/96.9	86.9/97.8/94.4	90.7/97.9/93.5	82.8/95.9/92.6	89.0/97.5/95.5	94.4/98.7/94.9	**97.9/99.5/97.7**
	Hazelnut	60.8/69.8/86.4	99.8/**100.**/99.3	99.9/99.9/99.3	98.8/99.2/98.6	99.5/99.7/97.3	**100./100./100.**	**100./100./100.**
	Metal Nut	**100./100.**/99.5	99.2/99.9/99.5	96.9/99.3/96.1	92.9/98.4/92.2	99.1/96.0/91.6	99.9/**100.**/99.5	**100./100./100.**
	Pill	97.5/99.6/96.8	93.7/98.7/95.7	88.2/97.7/92.5	77.1/94.4/91.7	95.7/98.5/94.5	97.0/99.5/96.2	**99.1/99.9/98.3**
	Screw	97.7/99.3/95.8	87.5/96.5/89.0	76.7/90.6/87.7	69.9/88.4/85.4	90.7/**99.7/97.9**	94.7/97.9/94.0	**98.4**/99.5/96.1
	Toothbrush	97.2/99.0/94.7	94.2/97.4/95.2	89.7/95.7/92.3	71.7/89.3/84.5	99.7/99.9/99.2	98.3/99.3/98.4	**100./100./100.**
	Transistor	94.2/95.2/90.0	99.8/98.0/93.8	99.2/98.7/97.6	78.2/79.5/68.8	99.8/99.6/97.4	**100./100./100.**	99.0/98.0/96.4
	Zipper	99.5/99.9/99.2	95.8/99.5/97.1	99.0/99.7/98.3	88.4/96.3/93.1	95.1/99.1/94.4	99.3/99.8/97.5	**100./100./100.**
Textures	Carpet	98.5/99.6/97.2	**99.8**/99.9/**99.4**	95.7/98.7/93.2	95.9/98.8/94.9	99.4/99.9/98.3	**99.8**/99.9/**99.4**	**99.8/100.**/98.9
	Grid	98.0/99.4/96.5	98.2/99.5/97.3	97.6/99.2/96.4	97.9/99.2/96.6	98.5/99.8/97.7	**100./100./100.**	99.9/**100.**/99.1
	Leather	**100./100./100.**	**100./100./100.**	**100./100./100.**	99.2/99.8/98.9	99.8/99.7/97.6	**100./100./100.**	**100./100./100.**
	Tile	98.3/99.3/96.4	99.3/99.8/98.2	99.3/99.8/98.8	97.0/98.9/95.3	96.8/99.9/98.4	98.2/99.3/95.4	**100./100./100.**
	Wood	99.2/99.8/98.3	98.6/99.6/96.6	98.4/99.5/96.7	99.9/**100.**/99.2	99.7/**100./100.**	98.8/99.6/96.6	99.8/99.9/99.2
	Mean	94.6/96.5/95.2	96.5/98.8/96.2	95.3/98.4/95.8	89.2/95.5/91.6	97.2/99.0/96.5	98.6/99.6/97.8	**99.6/99.8/99.0**

表A15. 在MVTec-AD数据集上使用AUROC/AP/F_1-max/AUPRO指标的多类别异常定位性能（按类别） s.

	Method →	RD4AD [10]	UniAD [60]	SimpleNet [34]	DeSTSeg [67]	DiAD [18]	MambaAD [17]	Dinomaly
	Category ↓	CVPR'22	NeurIPS'22	CVPR'23	CVPR'23	AAAI'24	Arxiv'24	Ours
Objects	Bottle	97.8/68.2/67.6/94.0	98.1/66.0/69.2/93.1	97.2/53.8/62.4/89.0	93.3/61.7/56.0/67.5	98.4/52.2/54.8/86.6	98.8/79.7/76.7/95.2	**99.2/88.6/84.2/96.6**
	Cable	85.1/26.3/33.6/75.1	97.3/39.9/45.2/86.1	96.7/42.4/51.2/85.4	89.3/37.5/40.5/49.4	96.8/50.1/57.8/80.5	95.8/42.2/48.1/90.3	**98.6/72.0/74.3/94.2**
	Capsule	**98.8**/43.4/50.0/94.8	98.5/42.7/46.5/92.1	98.5/35.4/44.3/84.5	95.8/47.9/48.9/62.1	97.1/42.0/45.3/87.2	98.4/43.9/47.7/92.6	98.7/**61.4/60.3/97.2**
	Hazelnut	97.9/36.2/51.6/92.7	98.1/55.2/56.8/94.1	98.4/44.6/51.4/87.4	98.2/65.8/61.6/84.5	98.3/79.2/**80.4**/91.5	99.0/63.6/64.4/95.7	**99.4/82.2**/76.4/**97.0**
	Metal Nut	94.8/55.5/66.4/91.9	62.7/14.6/29.2/81.8	**98.0/83.1**/79.4/85.2	84.2/42.0/22.8/53.0	97.3/30.0/38.3/90.6	96.7/74.5/79.1/93.7	96.9/78.6/**86.7/94.9**
	Pill	97.5/63.4/65.2/95.8	95.0/44.0/53.9/95.3	96.5/72.4/67.7/81.9	96.2/61.7/41.8/27.9	95.7/46.0/51.4/89.0	97.4/64.0/66.5/95.7	**97.8/76.4/71.6/97.3**
	Screw	99.4/40.2/44.6/96.8	98.3/28.7/37.6/95.2	96.5/15.9/23.2/84.0	93.8/19.9/25.3/47.3	97.9/**60.6/59.6**/95.0	99.5/49.8/50.9/97.1	**99.6**/60.2/**59.6/98.3**
	Toothbrush	**99.0**/53.6/58.8/92.0	98.4/34.9/45.7/87.9	98.4/46.9/52.5/87.4	96.2/52.9/58.8/30.9	**99.0**/78.7/72.8/95.0	**99.0**/48.5/59.2/91.7	98.9/**51.5/62.6/95.3**
	Transistor	85.9/42.3/45.2/74.7	**97.9**/59.5/**64.6/93.5**	95.8/58.2/56.0/83.2	73.6/38.4/39.2/43.9	95.1/15.6/31.7/90.0	96.5/69.4/67.1/87.0	93.2/**59.9**/58.5/77.0
	Zipper	98.5/53.9/60.3/94.1	96.8/40.1/49.9/92.6	97.9/53.4/54.6/90.7	97.3/64.7/59.2/66.9	96.2/60.7/60.0/91.6	98.4/60.4/61.7/94.3	**99.2/79.5/75.4/97.2**
Textures	Carpet	99.0/58.5/60.4/95.1	98.5/49.9/51.1/94.4	97.4/38.7/43.2/90.6	93.6/59.9/58.9/89.3	98.6/42.2/46.4/90.6	99.2/60.0/63.3/96.7	**99.3/68.7/71.1/97.6**
	Grid	96.5/23.0/28.4/97.0	63.1/10.7/11.9/92.9	96.8/20.5/27.6/88.6/	97.0/42.1/46.9/86.8	96.6/66.0/64.1/94.0	99.2/47.4/47.7/97.0	**99.4/55.3/57.7/97.2**
	Leather	99.3/38.0/45.1/97.4	98.8/32.9/34.4/96.8	98.7/28.5/32.9/92.7	**99.5/71.5/66.5**/91.1	98.8/56.1/62.3/91.3	99.4/50.3/53.3/98.7	99.4/52.2/55.0/**97.6**
	Tile	95.3/48.5/60.5/85.8	91.8/42.1/50.6/78.4	95.7/60.5/59.9/90.6	93.0/71.0/66.2/87.1	92.4/65.7/64.1/**90.7**	93.8/45.1/54.8/80.0	**98.1/80.1/75.7**/90.5
	Wood	95.3/47.8/51.0/90.0	93.2/37.2/41.5/86.7	91.4/34.8/39.7/76.3	95.9/**77.3/71.3**/83.4	93.3/43.3/43.5/**97.5**	94.4/46.2/48.2/91.2	**97.6**/72.8/68.4/94.0
	Mean	96.1/48.6/53.8/91.1	96.8/43.4/49.5/90.7	96.9/45.9/49.7/86.5	93.1/54.3/50.9/64.8	96.8/52.6/55.5/90.7	97.7/56.3/59.2/93.1	**98.4/69.3/69.2/94.8**

表A16. 在VisA数据集上使用AUROC/AP/F1-max指标进行多类别异常检测的各类别性能。

Method →	RD4AD [10]	UniAD [60]	SimpleNet [34]	DeSTSeg [67]	DiAD [18]	MambaAD	Dinomaly
Category ↓	CVPR'22	NeurIPS'22	CVPR'23	CVPR'23	AAAI'24	Arxiv'24	Ours
pcb1	96.2/95.5/91.9	92.8/92.7/87.8	91.6/91.9/86.0	87.6/83.1/83.7	88.1/88.7/80.7	95.4/93.0/91.6	**99.1/99.1/96.6**
pcb2	97.8/97.8/94.2	87.8/87.7/83.1	92.4/93.3/84.5	86.5/85.8/82.6	91.4/91.4/84.7	94.2/93.7/89.3	**99.3/99.2/97.0**
pcb3	96.4/96.2/91.0	78.6/78.6/76.1	89.1/91.1/82.6	93.7/95.1/87.0	86.2/87.6/77.6	93.7/94.1/86.7	**98.9/98.9/96.1**
pcb4	99.9/99.9/99.0	98.8/98.8/94.3	97.0/97.0/93.5	97.8/97.8/92.7	99.6/99.5/97.0	**99.9/99.9/98.5**	99.8/99.8/98.0
macaroni1	75.9/ 1.5/76.8	79.9/79.8/72.7	85.9/82.5/73.1	76.6/69.0/71.0	85.7/85.2/78.8	91.6/89.8/81.6	**98.0/97.6/94.2**
macaroni2	88.3/84.5/83.8	71.6/71.6/69.9	68.3/54.3/59.7	68.9/62.1/67.7	62.5/57.4/69.6	81.6/78.0/73.8	**95.9/95.7/90.7**
capsules	82.2/90.4/81.3	55.6/55.6/76.9	74.1/82.8/74.6	87.1/93.0/84.2	58.2/69.0/78.5	91.8/95.0/88.8	**98.6/99.0/97.1**
candle	92.3/92.9/86.0	94.1/94.0/86.1	84.1/73.3/76.6	94.9/94.8/89.2	92.8/92.0/87.6	96.8/96.9/90.1	**98.7/98.8/95.1**
cashew	92.0/95.8/90.7	92.8/92.8/91.4	88.0/91.3/84.7	92.0/96.1/88.1	91.5/95.7/89.7	94.5/97.3/91.1	**98.7/99.4/97.0**
chewinggum	94.9/97.5/92.1	96.3/96.2/95.2	96.4/98.2/93.8	95.8/98.3/94.7	99.1/99.5/95.9	97.7/98.9/94.2	**99.8/99.9/99.0**
fryum	95.3/97.9/91.5	83.0/83.0/85.0	88.4/93.0/83.3	92.1/96.1/89.5	89.8/95.0/87.2	95.2/97.7/90.5	**98.8/99.4/96.5**
pipe_fryum	97.9/98.9/96.5	94.7/94.7/93.9	90.8/95.5/88.6	94.1/97.1/91.9	96.2/98.1/93.7	98.7/99.3/97.0	**99.2/99.7/97.0**
Mean	92.4/92.4/89.6	85.5/85.5/84.4	87.2/87.0/81.8	88.9/89.0/85.2	86.8/88.3/85.1	94.3/94.5/89.4	**98.7/98.9/96.2**

Table A17. Per-class performance on **VisA** dataset for multi-class anomaly localization with AUROC/AP/F_1-max/AUPRO metrics.

Method → / Category ↓	RD4AD [10] CVPR'22	UniAD [60] NeurIPS'22	SimpleNet [34] CVPR'23	DeSTSeg [67] CVPR'23	DiAD [18] AAAI'24	MambaAD Arxiv'24	Dinomaly Ours
pcb1	99.4/66.2/62.4/**95.8**	93.3/ 3.9/ 8.3/64.1	99.2/86.1/78.8/83.6	95.8/46.4/49.0/83.2	98.7/49.6/52.8/80.2	**99.8**/77.1/72.4/92.8	99.5/**87.9/80.5**/95.1
pcb2	98.0/22.3/30.0/90.8	93.9/ 4.2/ 9.2/66.9	96.6/ 8.9/18.6/85.7	97.3/14.6/28.2/79.9	95.2/ 7.5/16.7/67.0	**98.9**/13.3/23.4/89.6	98.0/**47.0/49.8/91.3**
pcb3	97.9/26.2/35.2/93.9	97.3/13.8/21.9/70.6	97.2/31.0/36.1/85.1	97.7/28.1/33.4/62.4	96.7/ 8.0/18.8/68.9	**99.1**/18.3/27.4/89.1	98.4/**41.7/45.3/94.6**
pcb4	97.8/31.4/37.0/88.7	94.9/14.7/22.9/72.3	93.9/23.9/32.9/61.1	95.8/**53.0/53.2**/76.9	97.0/17.6/27.2/85.0	98.6/47.0/46.9/87.6	**98.7**/50.5/53.1/**94.4**
macaroni1	99.4/ 2.9/6.9/95.3	97.4/ 3.7/ 9.7/84.0	98.9/ 3.5/8.4/92.0	99.1/ 5.8/13.4/62.4	94.1/10.2/16.7/68.5	99.5/17.5/27.6/95.2	**99.6/33.5/40.6/96.4**
macaroni2	99.7/13.2/21.8/97.4	95.2/ 0.9/ 4.3/76.6	93.2/ 0.6/ 3.9/77.8	98.5/ 6.3/14.4/70.0	93.6/ 0.9/ 2.8/73.1	99.5/ 9.2/16.1/96.2	**99.7/24.7/36.1/98.7**
capsules	99.4/60.4/60.8/93.1	88.7/ 3.0/ 7.4/43.7	97.1/52.9/53.3/73.7	96.9/33.2/ 9.1/76.7	97.3/10.0/21.0/77.9	99.1/61.3/59.8/91.8	**99.6/65.0/66.6/97.4**
candle	99.1/25.3/35.8/94.9	98.5/17.6/27.9/91.6	97.6/ 8.4/16.5/87.6	98.7/39.9/45.8/69.0	97.3/12.8/22.8/89.4	99.0/23.2/32.4/**95.5**	**99.4/43.0/47.9**/95.4
cashew	91.7/44.2/49.7/86.2	98.6/51.7/58.3/87.9	**98.9/68.9/66.0**/84.1	87.9/47.6/52.1/66.3	90.9/53.1/60.9/61.8	94.3/46.8/51.4/87.8	97.1/64.5/62.4/**94.0**
chewinggum	98.7/59.9/61.7/76.9	98.8/54.9/56.1/81.3	97.9/26.8/29.8/78.3	98.8/**86.9/81.0**/68.3	94.7/11.9/25.8/59.5	98.1/57.5/59.9/79.7	**99.1**/65.0/67.7/**88.1**
fryum	97.0/47.6/51.5/93.4	95.9/34.0/40.6/76.2	93.0/39.1/45.4/85.1	88.1/35.2/38.5/47.7	**97.6/58.6/60.1**/81.3	96.9/47.8/51.9/91.6	96.6/51.6/53.4/**93.5**
pipe_fryum	99.1/56.8/58.8/**95.4**	98.9/50.2/57.7/91.5	98.5/65.6/63.4/83.0	98.9/78.8/72.7/45.9	**99.4/72.7/69.9**/89.9	99.1/53.5/58.5/95.1	99.2/64.3/65.1/95.2
Mean	98.1/38.0/42.6/91.8	95.9/21.0/27.0/75.6	96.8/34.7/37.8/81.4	96.1/39.6/43.4/67.4	96.0/26.1/33.0/75.2	98.5/39.4/44.0/91.0	**98.7/53.2/55.7/94.5**

Table A18. Per-class performance on **Real-IAD** dataset for multi-class anomaly detection with AUROC/AP/F_1-max metrics.

Method → / Category ↓	RD4AD [10] CVPR'22	UniAD [60] NeurIPS'22	SimpleNet [34] CVPR'23	DeSTSeg [67] CVPR'23	DiAD [18] AAAI'24	MambaAD Arxiv'24	Dinomaly Ours
audiojack	76.2/63.2/60.8	81.4/76.6/64.9	58.4/44.2/50.9	81.1/72.6/64.5	76.5/54.3/65.7	84.2/76.5/67.4	**86.8/82.4/72.2**
bottle cap	89.5/86.3/81.0	92.5/91.7/81.7	54.1/47.6/60.3	78.1/74.6/68.1	91.6/**94.0/87.9**	**92.8**/92.0/82.1	89.9/86.7/81.2
button battery	73.3/78.9/76.1	75.9/81.6/76.3	52.5/60.5/72.4	86.7/89.2/83.5	80.5/71.3/70.6	79.8/85.3/77.8	86.6/88.9/82.1
end cap	79.8/84.0/77.8	80.9/86.1/78.0	51.6/60.8/72.9	77.9/81.1/77.1	85.1/83.4/**84.8**	78.0/82.8/77.2	**87.0/87.5**/83.4
eraser	90.0/88.7/79.7	**90.3/89.2/80.2**	46.4/39.1/55.8	84.6/82.9/71.8	80.0/80.0/77.3	87.5/86.2/76.1	**90.3**/87.6/78.6
fire hood	78.3/70.1/64.5	80.6/74.8/66.4	58.1/41.9/54.4	81.7/72.4/67.7	83.3/**81.7/80.5**	79.3/72.5/64.8	**83.8**/76.2/69.5
mint	65.8/63.1/64.8	67.0/66.6/64.6	52.4/50.3/63.7	58.4/55.8/63.7	**76.7/76.7/76.0**	70.1/70.8/65.5	73.1/72.0/67.7
mounts	88.6/79.9/74.8	87.6/77.3/77.2	58.7/48.1/52.4	74.7/56.5/63.1	75.3/74.5/**82.5**	86.8/78.0/73.5	**90.4/84.2**/78.0
pcb	79.5/85.8/79.7	81.0/88.2/79.1	54.5/66.0/75.5	82.0/88.7/79.6	86.0/85.1/85.4	89.1/93.7/84.0	**92.0/95.3/87.0**
phone battery	87.5/83.3/77.1	83.6/80.0/71.6	51.6/43.8/58.0	83.3/81.8/72.1	82.3/77.7/75.9	90.2/88.9/80.5	**92.9/91.6/82.5**
plastic nut	80.3/68.0/64.4	80.0/69.2/63.7	59.2/40.3/51.8	83.1/75.4/66.5	71.9/58.2/65.6	87.1/80.7/70.7	**88.3/81.8/74.7**
plastic plug	81.9/74.3/68.8	81.4/75.9/67.6	48.2/38.4/54.6	71.7/63.1/60.0	88.7/**89.2/90.9**	85.7/82.2/72.6	**90.5**/86.4/78.6
porcelain doll	86.3/76.3/71.5	85.1/75.2/69.3	66.3/54.5/52.1	78.7/66.2/64.3	72.6/66.8/65.2	**88.0/82.2/74.1**	85.1/73.3/69.6
regulator	66.9/48.8/47.7	56.9/41.5/44.5	50.5/29.0/43.9	79.2/63.5/56.9	72.1/71.4/**78.2**	69.7/58.7/50.4	**85.2/78.9**/69.8
rolled strip base	97.5/98.7/94.7	98.7/99.3/96.5	59.0/75.7/79.8	96.5/98.2/93.0	68.4/55.9/56.8	98.0/99.0/95.0	**99.2/99.6/97.1**
sim card set	91.6/91.8/84.8	89.7/90.3/83.2	63.1/69.7/70.8	95.5/96.2/**89.2**	72.6/53.7/61.5	94.4/95.1/87.2	**95.8/96.3**/88.8
switch	84.3/87.2/77.9	85.5/88.6/78.4	62.2/66.8/68.6	90.1/92.8/83.1	73.4/49.4/61.2	91.7/94.0/85.4	**97.8/98.1/93.3**
tape	96.0/95.1/87.6	**97.2/96.2/89.4**	49.9/41.1/54.5	94.5/93.4/85.9	73.9/57.8/66.1	96.8/95.9/89.3	96.9/95.0/88.8
terminalblock	89.4/89.7/83.1	87.5/89.1/81.0	59.8/64.7/68.8	83.1/86.2/76.6	62.1/36.4/47.8	96.1/96.8/90.0	**96.7/97.4/91.1**
toothbrush	82.0/83.8/77.2	78.4/80.1/75.6	65.9/70.0/70.1	83.7/85.3/79.0	**91.2/93.7/90.9**	85.1/86.2/80.3	90.4/91.9/83.4
toy	69.4/74.2/75.9	68.4/75.1/74.8	57.8/64.4/73.4	70.3/74.8/75.4	66.2/57.3/59.8	83.0/87.5/79.6	85.6/89.1/81.9
toy brick	63.6/56.1/59.0	**77.0/71.1**/66.2	58.3/49.7/58.2	73.2/68.7/**63.3**	68.4/45.3/55.9	70.5/63.7/61.6	72.3/65.1/63.4
transistor1	91.0/94.0/85.1	93.7/95.9/88.9	62.2/69.2/72.1	90.2/92.1/84.6	73.1/63.1/62.7	94.4/96.0/89.0	**97.4/98.2/93.1**
u block	89.5/85.0/74.2	88.8/84.2/75.5	62.4/48.4/51.8	80.1/73.9/64.3	75.2/68.4/67.9	89.7/**85.7/75.3**	**89.9**/84.0/75.2
usb	84.9/84.3/75.1	78.7/79.4/69.1	57.0/55.3/62.9	87.8/88.0/78.3	58.9/37.4/45.7	**92.0/92.2/84.5**	**92.0**/91.6/83.3
usb adaptor	71.1/61.4/62.2	76.8/71.3/64.9	47.5/38.4/56.5	80.1/**74.9**/67.4	76.9/60.2/67.2	79.4/76.0/66.3	**81.5**/74.5/**69.4**
vcpill	85.1/80.3/72.4	87.1/84.0/74.7	59.0/48.7/56.4	83.8/81.5/69.9	64.1/40.4/56.2	88.3/87.7/77.4	**92.0/91.2/82.0**
wooden beads	81.2/78.9/70.9	78.4/77.2/67.8	55.1/52.0/60.2	82.4/78.5/73.0	62.1/56.4/65.9	82.5/81.7/71.8	**87.3/85.8/77.4**
woodstick	76.9/61.2/58.1	80.8/72.6/63.6	58.2/35.6/45.2	80.4/69.2/60.3	74.1/66.0/62.1	80.4/69.0/63.4	**84.0/73.3/65.6**
zipper	95.3/97.2/91.2	98.2/98.9/95.3	77.2/86.7/77.6	96.9/98.1/93.5	86.0/87.0/84.0	**99.2/99.6/96.9**	99.1/99.5/96.5
Mean	82.4/79.0/73.9	83.0/80.9/74.3	57.2/53.4/61.5	82.3/79.2/73.2	75.6/66.4/69.9	86.3/84.6/77.0	**89.3/86.8/80.2**

表 A17. 各类性能表现　　数控VisA数据集上进行多类别异常定位的AUROC/　　AP/F_1-max/AUPRO指标。

Method → / Category ↓	RD4AD [10] CVPR'22	UniAD [60] NeurIPS'22	SimpleNet [34] CVPR'23	DeSTSeg [67] CVPR'23	DiAD [18] AAAI'24	MambaAD Arxiv'24	Dinomaly Ours
pcb1	99.4/66.2/62.4/**95.8**	93.3/ 3.9/ 8.3/64.1	99.2/86.1/78.8/83.6	95.8/46.4/49.0/83.2	98.7/49.6/52.8/80.2	**99.8**/77.1/72.4/92.8	99.5/**87.9/80.5**/95.1
pcb2	98.0/22.3/30.0/90.8	93.9/ 4.2/ 9.2/66.9	96.6/ 8.9/18.6/85.7	97.3/14.6/28.2/79.9	95.2/ 7.5/16.7/67.0	**98.9**/13.3/23.4/89.6	98.0/**47.0/49.8/91.3**
pcb3	97.9/26.2/35.2/93.9	97.3/13.8/21.9/70.6	97.2/31.0/36.1/85.1	97.7/28.1/33.4/62.4	96.7/ 8.0/18.8/68.9	**99.1**/18.3/27.4/89.1	98.4/**41.7/45.3/94.6**
pcb4	97.8/31.4/37.0/88.7	94.9/14.7/22.9/72.3	93.9/23.9/32.9/61.1	95.8/**53.0/53.2**/76.9	97.0/17.6/27.2/85.0	98.6/47.0/46.9/87.6	**98.7**/50.5/53.1/**94.4**
macaroni1	99.4/ 2.9/6.9/95.3	97.4/ 3.7/ 9.7/84.0	98.9/ 3.5/8.4/92.0	99.1/ 5.8/13.4/62.4	94.1/10.2/16.7/68.5	99.5/17.5/27.6/95.2	**99.6/33.5/40.6/96.4**
macaroni2	99.7/13.2/21.8/97.4	95.2/ 0.9/ 4.3/76.6	93.2/ 0.6/ 3.9/77.8	98.5/ 6.3/14.4/70.0	93.6/ 0.9/ 2.8/73.1	99.5/ 9.2/16.1/96.2	**99.7/24.7/36.1/98.7**
capsules	99.4/60.4/60.8/93.1	88.7/ 3.0/ 7.4/43.7	97.1/52.9/53.3/73.7	96.9/33.2/ 9.1/76.7	97.3/10.0/21.0/77.9	99.1/61.3/59.8/91.8	**99.6/65.0/66.6/97.4**
candle	99.1/25.3/35.8/94.9	98.5/17.6/27.9/91.6	97.6/ 8.4/16.5/87.6	98.7/39.9/45.8/69.0	97.3/12.8/22.8/89.4	99.0/23.2/32.4/**95.5**	**99.4/43.0/47.9**/95.4
cashew	91.7/44.2/49.7/86.2	98.6/51.7/58.3/87.9	**98.9/68.9/66.0**/84.1	87.9/47.6/52.1/66.3	90.9/53.1/60.9/61.8	94.3/46.8/51.4/87.8	97.1/64.5/62.4/**94.0**
chewinggum	98.7/59.9/61.7/76.9	98.8/54.9/56.1/81.3	97.9/26.8/29.8/78.3	98.8/**86.9/81.0**/68.3	94.7/11.9/25.8/59.5	98.1/57.5/59.9/79.7	**99.1**/65.0/67.7/**88.1**
fryum	97.0/47.6/51.5/93.4	95.9/34.0/40.6/76.2	93.0/39.1/45.4/85.1	88.1/35.2/38.5/47.7	**97.6/58.6/60.1**/81.3	96.9/47.8/51.9/91.6	96.6/51.6/53.4/**93.5**
pipe_fryum	99.1/56.8/58.8/**95.4**	98.9/50.2/57.7/91.5	98.5/65.6/63.4/83.0	98.9/78.8/72.7/45.9	**99.4/72.7/69.9**/89.9	99.1/53.5/58.5/95.1	99.2/64.3/65.1/95.2
Mean	98.1/38.0/42.6/91.8	95.9/21.0/27.0/75.6	96.8/34.7/37.8/81.4	96.1/39.6/43.4/67.4	96.0/26.1/33.0/75.2	98.5/39.4/44.0/91.0	**98.7/53.2/55.7/94.5**

表A18. 在Real-IAD数据集上使用AUROC/AP/F_1-max指标进行多类别异常检测的各类别性能。

Method → / Category ↓	RD4AD [10] CVPR'22	UniAD [60] NeurIPS'22	SimpleNet [34] CVPR'23	DeSTSeg [67] CVPR'23	DiAD [18] AAAI'24	MambaAD Arxiv'24	Dinomaly Ours
audiojack	76.2/63.2/60.8	81.4/76.6/64.9	58.4/44.2/50.9	81.1/72.6/64.5	76.5/54.3/65.7	84.2/76.5/67.4	**86.8/82.4/72.2**
bottle cap	89.5/86.3/81.0	92.5/91.7/81.7	54.1/47.6/60.3	78.1/74.6/68.1	91.6/**94.0/87.9**	**92.8**/92.0/82.1	89.9/86.7/81.2
button battery	73.3/78.9/76.1	75.9/81.6/76.3	52.5/60.5/72.4	86.7/89.2/83.5	80.5/71.3/70.6	79.8/85.3/77.8	86.6/88.9/82.1
end cap	79.8/84.0/77.8	80.9/86.1/78.0	51.6/60.8/72.9	77.9/81.1/77.1	85.1/83.4/**84.8**	78.0/82.8/77.2	**87.0/87.5**/83.4
eraser	90.0/88.7/79.7	**90.3/89.2/80.2**	46.4/39.1/55.8	84.6/82.9/71.8	80.0/80.0/77.3	87.5/86.2/76.1	**90.3**/87.6/78.6
fire hood	78.3/70.1/64.5	80.6/74.8/66.4	58.1/41.9/54.4	81.7/72.4/67.7	83.3/**81.7/80.5**	79.3/72.5/64.8	**83.8**/76.2/69.5
mint	65.8/63.1/64.8	67.0/66.6/64.6	52.4/50.3/63.7	58.4/55.8/63.7	**76.7/76.7/76.0**	70.1/70.8/65.5	73.1/72.0/67.7
mounts	88.6/79.9/74.8	87.6/77.3/77.2	58.7/48.1/52.4	74.7/56.5/63.1	75.3/74.5/**82.5**	86.8/78.0/73.5	**90.4/84.2**/78.0
pcb	79.5/85.8/79.7	81.0/88.2/79.1	54.5/66.0/75.5	82.0/88.7/79.6	86.0/85.1/85.4	89.1/93.7/84.0	**92.0/95.3/87.0**
phone battery	87.5/83.3/77.1	83.6/80.0/71.6	51.6/43.8/58.0	83.3/81.8/72.1	82.3/77.7/75.9	90.2/88.9/80.5	**92.9/91.6/82.5**
plastic nut	80.3/68.0/64.4	80.0/69.2/63.7	59.2/40.3/51.8	83.1/75.4/66.5	71.9/58.2/65.6	87.1/80.7/70.7	**88.3/81.8/74.7**
plastic plug	81.9/74.3/68.8	81.4/75.9/67.6	48.2/38.4/54.6	71.7/63.1/60.0	88.7/**89.2/90.9**	85.7/82.2/72.6	**90.5**/86.4/78.6
porcelain doll	86.3/76.3/71.5	85.1/75.2/69.3	66.3/54.5/52.1	78.7/66.2/64.3	72.6/66.8/65.2	**88.0/82.2/74.1**	85.1/73.3/69.6
regulator	66.9/48.8/47.7	56.9/41.5/44.5	50.5/29.0/43.9	79.2/63.5/56.9	72.1/71.4/**78.2**	69.7/58.7/50.4	**85.2/78.9**/69.8
rolled strip base	97.5/98.7/94.7	98.7/99.3/96.5	59.0/75.7/79.8	96.5/98.2/93.0	68.4/55.9/56.8	98.0/99.0/95.0	**99.2/99.6/97.1**
sim card set	91.6/91.8/84.8	89.7/90.3/83.2	63.1/69.7/70.8	95.5/96.2/**89.2**	72.6/53.7/61.5	94.4/95.1/87.2	**95.8/96.3**/88.8
switch	84.3/87.2/77.9	85.5/88.6/78.4	62.2/66.8/68.6	90.1/92.8/83.1	73.4/49.4/61.2	91.7/94.0/85.4	**97.8/98.1/93.3**
tape	96.0/95.1/87.6	**97.2/96.2/89.4**	49.9/41.1/54.5	94.5/93.4/85.9	73.9/57.8/66.1	96.8/95.9/89.3	96.9/95.0/88.8
terminalblock	89.4/89.7/83.1	87.5/89.1/81.0	59.8/64.7/68.8	83.1/86.2/76.6	62.1/36.4/47.8	96.1/96.8/90.0	**96.7/97.4/91.1**
toothbrush	82.0/83.8/77.2	78.4/80.1/75.6	65.9/70.0/70.1	83.7/85.3/79.0	**91.2/93.7/90.9**	85.1/86.2/80.3	90.4/91.9/83.4
toy	69.4/74.2/75.9	68.4/75.1/74.8	57.8/64.4/73.4	70.3/74.8/75.4	66.2/57.3/59.8	83.0/87.5/79.6	85.6/89.1/81.9
toy brick	63.6/56.1/59.0	**77.0/71.1**/66.2	58.3/49.7/58.2	73.2/68.7/**63.3**	68.4/45.3/55.9	70.5/63.7/61.6	72.3/65.1/63.4
transistor1	91.0/94.0/85.1	93.7/95.9/88.9	62.2/69.2/72.1	90.2/92.1/84.6	73.1/63.1/62.7	94.4/96.0/89.0	**97.4/98.2/93.1**
u block	89.5/85.0/74.2	88.8/84.2/75.5	62.4/48.4/51.8	80.1/73.9/64.3	75.2/68.4/67.9	89.7/**85.7/75.3**	**89.9**/84.0/75.2
usb	84.9/84.3/75.1	78.7/79.4/69.1	57.0/55.3/62.9	87.8/88.0/78.3	58.9/37.4/45.7	**92.0/92.2/84.5**	**92.0**/91.6/83.3
usb adaptor	71.1/61.4/62.2	76.8/71.3/64.9	47.5/38.4/56.5	80.1/**74.9**/67.4	76.9/60.2/67.2	79.4/76.0/66.3	**81.5**/74.5/**69.4**
vcpill	85.1/80.3/72.4	87.1/84.0/74.7	59.0/48.7/56.4	83.8/81.5/69.9	64.1/40.4/56.2	88.3/87.7/77.4	**92.0/91.2/82.0**
wooden beads	81.2/78.9/70.9	78.4/77.2/67.8	55.1/52.0/60.2	82.4/78.5/73.0	62.1/56.4/65.9	82.5/81.7/71.8	**87.3/85.8/77.4**
woodstick	76.9/61.2/58.1	80.8/72.6/63.6	58.2/35.6/45.2	80.4/69.2/60.3	74.1/66.0/62.1	80.4/69.0/63.4	**84.0/73.3/65.6**
zipper	95.3/97.2/91.2	98.2/98.9/95.3	77.2/86.7/77.6	96.9/98.1/93.5	86.0/87.0/84.0	**99.2/99.6/96.9**	99.1/99.5/96.5
Mean	82.4/79.0/73.9	83.0/80.9/74.3	57.2/53.4/61.5	82.3/79.2/73.2	75.6/66.4/69.9	86.3/84.6/77.0	**89.3/86.8/80.2**

Table A19. Per-class performance on **Real-IAD** dataset for multi-class anomaly localization with AUROC/AP/F_1-max/AUPRO metrics.

Method →	RD4AD [10]	UniAD [60]	SimpleNet [34]	DeSTSeg [67]	DiAD [18]	MambaAD [17]	Dinomaly
Category ↓	CVPR'22	NeurIPS'22	CVPR'23	CVPR'23	AAAI'24	Arxiv'24	Ours
audiojack	96.6/12.8/22.1/79.6	97.6/20.0/31.0/83.7	74.4/ 0.9/ 4.8/38.0	95.5/25.4/31.9/52.6	91.6/ 1.0/ 3.9/63.3	97.7/21.6/29.5/83.9	**98.7/48.1/54.5/91.7**
bottle cap	99.5/18.9/29.9/95.7	99.5/19.4/29.6/96.0	85.3/ 2.3/ 5.7/45.1	94.5/25.3/31.1/25.3	94.6/ 4.9/11.4/73.0	**99.7**/30.6/34.6/97.2	**99.7/32.4/36.7/98.1**
button battery	97.6/33.8/37.8/86.5	96.7/28.5/34.4/77.5	75.9/ 3.2/ 6.6/40.5	98.3/**63.9/60.4**/36.9	84.1/ 1.4/ 5.3/66.9	98.1/46.7/49.5/86.2	**99.1**/46.9/56.7/**92.9**
end cap	96.7/12.5/22.5/89.2	95.8/ 8.8/17.4/85.4	63.1/ 0.5/ 2.8/25.7	89.6/14.4/22.7/29.5	81.3/ 2.0/ 6.9/38.2	97.0/12.0/19.6/89.4	**99.1/26.2/32.9/96.0**
eraser	**99.5**/30.8/36.7/96.0	99.3/24.4/30.9/94.1	80.6/ 2.7/ 7.1/42.8	95.8/52.7/53.9/46.7	91.1/ 7.7/15.4/67.5	99.2/30.2/38.3/93.7	**99.5/39.6/43.3/96.4**
fire hood	98.9/27.7/35.2/87.9	98.6/23.4/32.2/85.3	70.5/ 0.3/ 2.2/25.3	97.3/27.1/35.3/34.7	91.8/ 3.2/ 9.2/66.7	98.7/25.1/31.3/86.3	**99.3/38.4/42.7/93.0**
mint	95.0/11.7/23.0/72.3	94.4/ 7.7/18.1/62.3	79.9/ 0.9/ 3.6/43.3	84.1/10.3/22.4/ 9.9	91.1/ 5.7/11.6/64.2	96.5/15.9/27.0/72.6	**96.9/22.0/32.5/77.6**
mounts	99.3/30.6/37.1/94.9	**99.4**/28.0/32.8/95.2	80.5/ 2.2/ 6.8/46.1	94.2/30.0/41.3/43.3	84.3/ 0.4/ 1.1/48.8	99.2/31.4/35.4/93.5	**99.4/39.9/44.3/95.6**
pcb	97.5/15.8/24.3/88.3	97.0/18.5/28.1/81.6	78.0/ 1.4/ 4.3/41.3	97.2/37.1/40.4/48.8	92.0/ 3.7/ 7.4/66.5	99.2/46.3/50.4/93.1	**99.3/55.0/56.3/95.7**
phone battery	77.3/22.6/31.7/94.5	85.5/11.2/21.6/88.5	43.4/ 0.1/ 0.9/11.8	79.5/25.6/33.8/39.5	96.8/ 5.3/11.4/85.4	99.4/36.3/41.3/95.3	**99.7/51.6/54.2/96.8**
phone battery	77.3/22.6/31.7/94.5	85.5/11.2/21.6/88.5	43.4/ 0.1/ 0.9/11.8	79.5/25.6/33.8/39.5	96.8/5.3/11.4/85.4	99.4/36.3/41.3/95.3	**99.7/51.6/54.2/96.8**
plastic nut	98.8/21.1/29.6/91.0	98.4/20.6/27.1/88.9	77.4/ 0.6/ 3.6/41.5	96.5/44.8/45.7/38.4	81.1/ 0.4/ 3.4/38.6	99.4/33.1/37.3/96.1	**99.7/41.0/45.0/97.4**
plastic plug	99.1/20.5/28.4/94.9	98.6/17.4/26.1/90.3	78.6/ 0.7/ 1.9/38.8	91.9/20.1/27.3/21.0	92.9/ 8.7/15.0/66.1	99.0/24.2/31.7/91.5	**99.4/31.7/37.2/96.4**
porcelain doll	99.2/24.8/34.6/95.7	98.7/14.1/24.5/93.2	81.8/ 2.0/ 6.4/47.0	93.1/35.9/40.3/24.8	93.1/ 1.4/ 4.8/70.4	99.2/**31.3/36.6**/95.4	**99.3**/27.9/33.9/**96.0**
regulator	98.0/7.8/16.1/88.6	95.5/9.1/17.4/76.1	76.6/0.1/0.6/38.1	88.8/18.9/23.6/17.5	84.2/0.4/1.5/44.4	97.6/20.6/29.8/87.0	**99.3/42.2/48.9/95.6**
rolled strip base	**99.7**/31.4/39.9/98.4	99.6/20.7/32.2/97.8	80.5/ 1.7/ 5.1/52.1	99.2/**48.7/50.1**/55.5	87.7/ 0.6/ 3.2/63.4	99.7/37.4/42.5/98.8	**99.7**/41.6/45.5/**98.5**
sim card set	98.5/40.2/44.2/89.5	97.9/31.6/39.8/85.0	71.0/ 6.8/14.3/30.8	**99.1/65.5/62.1**/73.9	89.9/ 1.7/ 5.8/60.4	98.8/51.1/50.6/89.4	99.0/52.1/52.9/**90.9**
switch	94.4/18.9/26.6/90.9	98.1/33.8/40.6/90.7	71.7/ 3.7/ 9.3/44.2	97.4/57.6/55.6/44.7	90.5/ 1.4/ 5.3/64.2	**98.2**/39.9/45.4/92.9	96.7/**62.3/63.6/95.9**
tape	99.7/42.4/47.8/98.4	99.7/29.2/36.9/97.5	77.5/ 1.2/ 3.9/41.4	99.0/61.7/57.6/48.2	81.7/ 0.4/ 2.7/47.3	**99.8**/47.1/48.2/98.0	**99.8/54.0/55.8**/98.8
terminalblock	99.5/27.4/35.8/97.6	99.2/23.1/30.5/94.4	87.0/ 0.8/ 3.6/54.8	96.6/40.6/44.1/34.8	75.5/ 0.1/ 1.1/38.5	**99.8**/35.3/39.7/98.2	**99.8/48.0/50.7/98.8**
toothbrush	96.9/26.1/34.2/88.7	95.7/16.4/25.3/84.3	84.7/ 7.2/14.8/52.6	94.3/30.0/37.3/42.8	82.0/ 1.9/ 6.6/54.5	**97.5**/27.8/36.7/**91.4**	96.9/**38.3/43.9**/90.4
toy	95.2/ 5.1/12.8/82.3	93.4/ 4.6/12.4/70.5	67.7/ 0.1/ 0.4/25.0	86.3/ 8.1/15.9/16.4	82.1/ 1.1/ 4.2/50.3	**96.0**/16.4/25.8/86.3	94.9/**22.5/32.1/91.0**
toy brick	96.4/16.0/24.6/75.3	**97.4**/17.1/27.6/**81.3**	86.5/ 5.2/11.1/56.3	94.7/24.6/30.8/45.5	93.5/ 3.1/ 8.1/66.4	96.6/18.0/25.8/74.7	96.8/**27.9/34.0**/76.6
transistor1	99.1/29.6/35.5/95.1	98.9/25.6/33.2/94.3	71.7/ 5.1/11.3/35.3	97.3/43.8/44.5/45.4	88.6/ 7.2/15.3/58.1	99.4/39.4/40.0/96.5	**99.6/53.5/53.3/97.8**
u block	99.6/40.5/45.2/96.9	99.3/22.3/29.6/94.3	76.2/ 4.8/12.2/34.0	96.9/**57.1/55.7**/38.5	88.8/ 1.6/ 5.4/54.2	**99.5**/37.8/46.1/95.4	**99.5**/41.8/45.6/**96.8**
usb	98.1/26.4/35.2/91.0	97.9/20.6/31.7/85.3	81.1/ 1.5/ 4.9/52.4	98.4/42.2/47.7/57.1	78.0/ 1.0/ 3.1/28.0	**99.2**/39.1/44.4/95.2	**99.2/45.0/48.7/97.5**
usb adaptor	94.5/ 9.8/17.9/73.1	96.6/10.5/19.0/78.4	67.9/ 0.2/ 1.3/28.9	94.9/**25.5/34.9**/36.4	94.0/ 2.3/ 6.6/75.5	97.3/15.3/22.6/82.5	**98.7**/23.7/32.7/**91.0**
vcpill	98.3/43.1/48.6/88.7	99.1/40.7/43.0/91.3	68.2/ 1.1/ 3.3/22.0	97.1/64.7/62.3/42.3	90.2/ 1.3/ 5.2/60.8	98.7/50.2/54.5/89.3	**99.1/66.4/66.7/93.7**
wooden beads	98.0/27.1/34.7/85.7	97.6/16.5/23.6/84.6	68.1/ 2.4/ 6.0/28.3	94.7/38.9/42.9/39.4	85.0/ 1.1/ 4.7/45.6	98.0/32.6/39.8/84.5	**99.1/45.8/50.1/90.5**
woodstick	97.8/30.7/38.4/85.0	94.0/36.2/44.3/77.2	76.1/ 1.4/ 6.0/32.0	97.9/**60.3/60.0**/51.0	90.9/ 2.6/ 8.0/60.7	97.7/40.1/44.9/82.7	**99.0**/50.9/52.1/**90.4**
zipper	99.1/44.7/50.2/96.3	98.4/32.5/36.1/95.1	89.9/23.3/31.2/55.5	98.2/35.3/39.0/78.5	90.2/12.5/18.8/53.5	99.3/58.2/61.3/97.6	**99.3/67.2/66.5/97.8**
Mean	97.3/25.0/32.7/89.6	97.3/21.1/29.2/86.7	75.7/ 2.8/ 6.5/39.0	94.6/37.9/41.7/40.6	88.0/ 2.9/ 7.1/58.1	98.5/33.0/38.7/90.5	**98.8/42.8/47.1/93.9**

表A19. 在Real-IAD数据集上使用AUROC/AP/F_1-max/AUPRO指标进行多类别异常定位的逐类别性能 。

Method →	RD4AD [10]	UniAD [60]	SimpleNet [34]	DeSTSeg [67]	DiAD [18]	MambaAD [17]	Dinomaly
Category ↓	CVPR'22	NeurIPS'22	CVPR'23	CVPR'23	AAAI'24	Arxiv'24	Ours
audiojack	96.6/12.8/22.1/79.6	97.6/20.0/31.0/83.7	74.4/ 0.9/ 4.8/38.0	95.5/25.4/31.9/52.6	91.6/ 1.0/ 3.9/63.3	97.7/21.6/29.5/83.9	**98.7/48.1/54.5/91.7**
bottle cap	99.5/18.9/29.9/95.7	99.5/19.4/29.6/96.0	85.3/ 2.3/ 5.7/45.1	94.5/25.3/31.1/25.3	94.6/ 4.9/11.4/73.0	**99.7**/30.6/34.6/97.2	**99.7/32.4/36.7/98.1**
button battery	97.6/33.8/37.8/86.5	96.7/28.5/34.4/77.5	75.9/ 3.2/ 6.6/40.5	98.3/**63.9/60.4**/36.9	84.1/ 1.4/ 5.3/66.9	98.1/46.7/49.5/86.2	**99.1**/46.9/56.7/**92.9**
end cap	96.7/12.5/22.5/89.2	95.8/ 8.8/17.4/85.4	63.1/ 0.5/ 2.8/25.7	89.6/14.4/22.7/29.5	81.3/ 2.0/ 6.9/38.2	97.0/12.0/19.6/89.4	**99.1/26.2/32.9/96.0**
eraser	**99.5**/30.8/36.7/96.0	99.3/24.4/30.9/94.1	80.6/ 2.7/ 7.1/42.8	95.8/52.7/53.9/46.7	91.1/ 7.7/15.4/67.5	99.2/30.2/38.3/93.7	**99.5/39.6/43.3/96.4**
fire hood	98.9/27.7/35.2/87.9	98.6/23.4/32.2/85.3	70.5/ 0.3/ 2.2/25.3	97.3/27.1/35.3/34.7	91.8/ 3.2/ 9.2/66.7	98.7/25.1/31.3/86.3	**99.3/38.4/42.7/93.0**
mint	95.0/11.7/23.0/72.3	94.4/ 7.7/18.1/62.3	79.9/ 0.9/ 3.6/43.3	84.1/10.3/22.4/ 9.9	91.1/ 5.7/11.6/64.2	96.5/15.9/27.0/72.6	**96.9/22.0/32.5/77.6**
mounts	99.3/30.6/37.1/94.9	**99.4**/28.0/32.8/95.2	80.5/ 2.2/ 6.8/46.1	94.2/30.0/41.3/43.3	84.3/ 0.4/ 1.1/48.8	99.2/31.4/35.4/93.5	**99.4/39.9/44.3/95.6**
pcb	97.5/15.8/24.3/88.3	97.0/18.5/28.1/81.6	78.0/ 1.4/ 4.3/41.3	97.2/37.1/40.4/48.8	92.0/ 3.7/ 7.4/66.5	99.2/46.3/50.4/93.1	**99.3/55.0/56.3/95.7**
phone battery	77.3/22.6/31.7/94.5	85.5/11.2/21.6/88.5	43.4/ 0.1/ 0.9/11.8	79.5/25.6/33.8/39.5	96.8/ 5.3/11.4/85.4	99.4/36.3/41.3/95.3	**99.7/51.6/54.2/96.8**
phone battery	77.3/22.6/31.7/94.5	85.5/11.2/21.6/88.5	43.4/ 0.1/ 0.9/11.8	79.5/25.6/33.8/39.5	96.8/5.3/11.4/85.4	99.4/36.3/41.3/95.3	**99.7/51.6/54.2/96.8**
plastic nut	98.8/21.1/29.6/91.0	98.4/20.6/27.1/88.9	77.4/ 0.6/ 3.6/41.5	96.5/44.8/45.7/38.4	81.1/ 0.4/ 3.4/38.6	99.4/33.1/37.3/96.1	**99.7/41.0/45.0/97.4**
plastic plug	99.1/20.5/28.4/94.9	98.6/17.4/26.1/90.3	78.6/ 0.7/ 1.9/38.8	91.9/20.1/27.3/21.0	92.9/ 8.7/15.0/66.1	99.0/24.2/31.7/91.5	**99.4/31.7/37.2/96.4**
porcelain doll	99.2/24.8/34.6/95.7	98.7/14.1/24.5/93.2	81.8/ 2.0/ 6.4/47.0	93.1/35.9/40.3/24.8	93.1/ 1.4/ 4.8/70.4	99.2/**31.3/36.6**/95.4	**99.3**/27.9/33.9/**96.0**
regulator	98.0/7.8/16.1/88.6	95.5/9.1/17.4/76.1	76.6/0.1/0.6/38.1	88.8/18.9/23.6/17.5	84.2/0.4/1.5/44.4	97.6/20.6/29.8/87.0	**99.3/42.2/48.9/95.6**
rolled strip base	**99.7**/31.4/39.9/98.4	99.6/20.7/32.2/97.8	80.5/ 1.7/ 5.1/52.1	99.2/**48.7/50.1**/55.5	87.7/ 0.6/ 3.2/63.4	99.7/37.4/42.5/98.8	**99.7**/41.6/45.5/**98.5**
sim card set	98.5/40.2/44.2/89.5	97.9/31.6/39.8/85.0	71.0/ 6.8/14.3/30.8	**99.1/65.5/62.1**/73.9	89.9/ 1.7/ 5.8/60.4	98.8/51.1/50.6/89.4	99.0/52.1/52.9/**90.9**
switch	94.4/18.9/26.6/90.9	98.1/33.8/40.6/90.7	71.7/ 3.7/ 9.3/44.2	97.4/57.6/55.6/44.7	90.5/ 1.4/ 5.3/64.2	**98.2**/39.9/45.4/92.9	96.7/**62.3/63.6/95.9**
tape	99.7/42.4/47.8/98.4	99.7/29.2/36.9/97.5	77.5/ 1.2/ 3.9/41.4	99.0/61.7/57.6/48.2	81.7/ 0.4/ 2.7/47.3	**99.8**/47.1/48.2/98.0	**99.8/54.0/55.8**/98.8
terminalblock	99.5/27.4/35.8/97.6	99.2/23.1/30.5/94.4	87.0/ 0.8/ 3.6/54.8	96.6/40.6/44.1/34.8	75.5/ 0.1/ 1.1/38.5	**99.8**/35.3/39.7/98.2	**99.8/48.0/50.7/98.8**
toothbrush	96.9/26.1/34.2/88.7	95.7/16.4/25.3/84.3	84.7/ 7.2/14.8/52.6	94.3/30.0/37.3/42.8	82.0/ 1.9/ 6.6/54.5	**97.5**/27.8/36.7/**91.4**	96.9/**38.3/43.9**/90.4
toy	95.2/ 5.1/12.8/82.3	93.4/ 4.6/12.4/70.5	67.7/ 0.1/ 0.4/25.0	86.3/ 8.1/15.9/16.4	82.1/ 1.1/ 4.2/50.3	**96.0**/16.4/25.8/86.3	94.9/**22.5/32.1/91.0**
toy brick	96.4/16.0/24.6/75.3	**97.4**/17.1/27.6/**81.3**	86.5/ 5.2/11.1/56.3	94.7/24.6/30.8/45.5	93.5/ 3.1/ 8.1/66.4	96.6/18.0/25.8/74.7	96.8/**27.9/34.0**/76.6
transistor1	99.1/29.6/35.5/95.1	98.9/25.6/33.2/94.3	71.7/ 5.1/11.3/35.3	97.3/43.8/44.5/45.4	88.6/ 7.2/15.3/58.1	99.4/39.4/40.0/96.5	**99.6/53.5/53.3/97.8**
u block	99.6/40.5/45.2/96.9	99.3/22.3/29.6/94.3	76.2/ 4.8/12.2/34.0	96.9/**57.1/55.7**/38.5	88.8/ 1.6/ 5.4/54.2	**99.5**/37.8/46.1/95.4	**99.5**/41.8/45.6/**96.8**
usb	98.1/26.4/35.2/91.0	97.9/20.6/31.7/85.3	81.1/ 1.5/ 4.9/52.4	98.4/42.2/47.7/57.1	78.0/ 1.0/ 3.1/28.0	**99.2**/39.1/44.4/95.2	**99.2/45.0/48.7/97.5**
usb adaptor	94.5/ 9.8/17.9/73.1	96.6/10.5/19.0/78.4	67.9/ 0.2/ 1.3/28.9	94.9/**25.5/34.9**/36.4	94.0/ 2.3/ 6.6/75.5	97.3/15.3/22.6/82.5	**98.7**/23.7/32.7/**91.0**
vcpill	98.3/43.1/48.6/88.7	99.1/40.7/43.0/91.3	68.2/ 1.1/ 3.3/22.0	97.1/64.7/62.3/42.3	90.2/ 1.3/ 5.2/60.8	98.7/50.2/54.5/89.3	**99.1/66.4/66.7/93.7**
wooden beads	98.0/27.1/34.7/85.7	97.6/16.5/23.6/84.6	68.1/ 2.4/ 6.0/28.3	94.7/38.9/42.9/39.4	85.0/ 1.1/ 4.7/45.6	98.0/32.6/39.8/84.5	**99.1/45.8/50.1/90.5**
woodstick	97.8/30.7/38.4/85.0	94.0/36.2/44.3/77.2	76.1/ 1.4/ 6.0/32.0	97.9/**60.3/60.0**/51.0	90.9/ 2.6/ 8.0/60.7	97.7/40.1/44.9/82.7	**99.0**/50.9/52.1/**90.4**
zipper	99.1/44.7/50.2/96.3	98.4/32.5/36.1/95.1	89.9/23.3/31.2/55.5	98.2/35.3/39.0/78.5	90.2/12.5/18.8/53.5	99.3/58.2/61.3/97.6	**99.3/67.2/66.5/97.8**
Mean	97.3/25.0/32.7/89.6	97.3/21.1/29.2/86.7	75.7/ 2.8/ 6.5/39.0	94.6/37.9/41.7/40.6	88.0/ 2.9/ 7.1/58.1	98.5/33.0/38.7/90.5	**98.8/42.8/47.1/93.9**

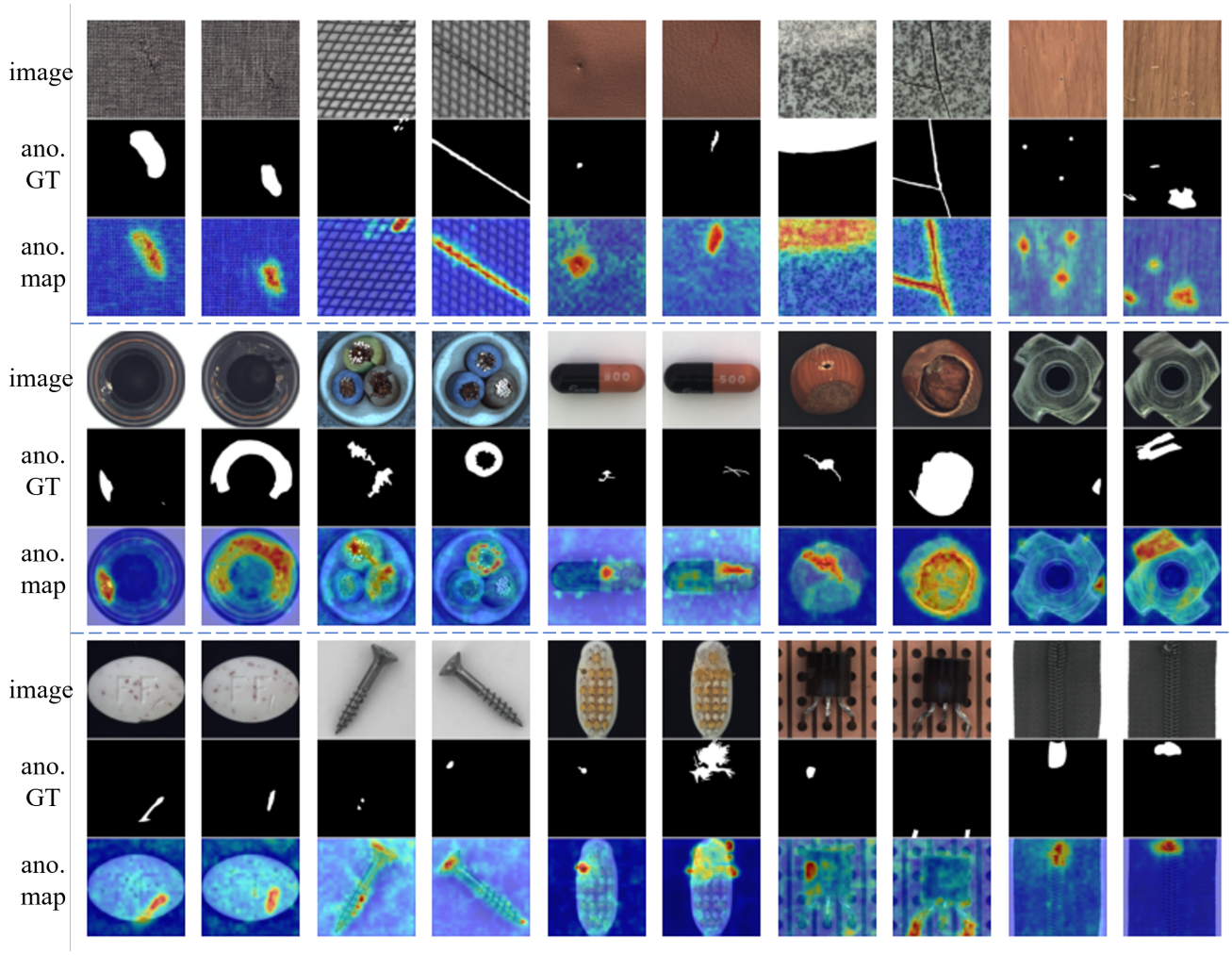


Figure A1. Anomaly maps visualization on MVTec-AD. All samples are randomly chosen.

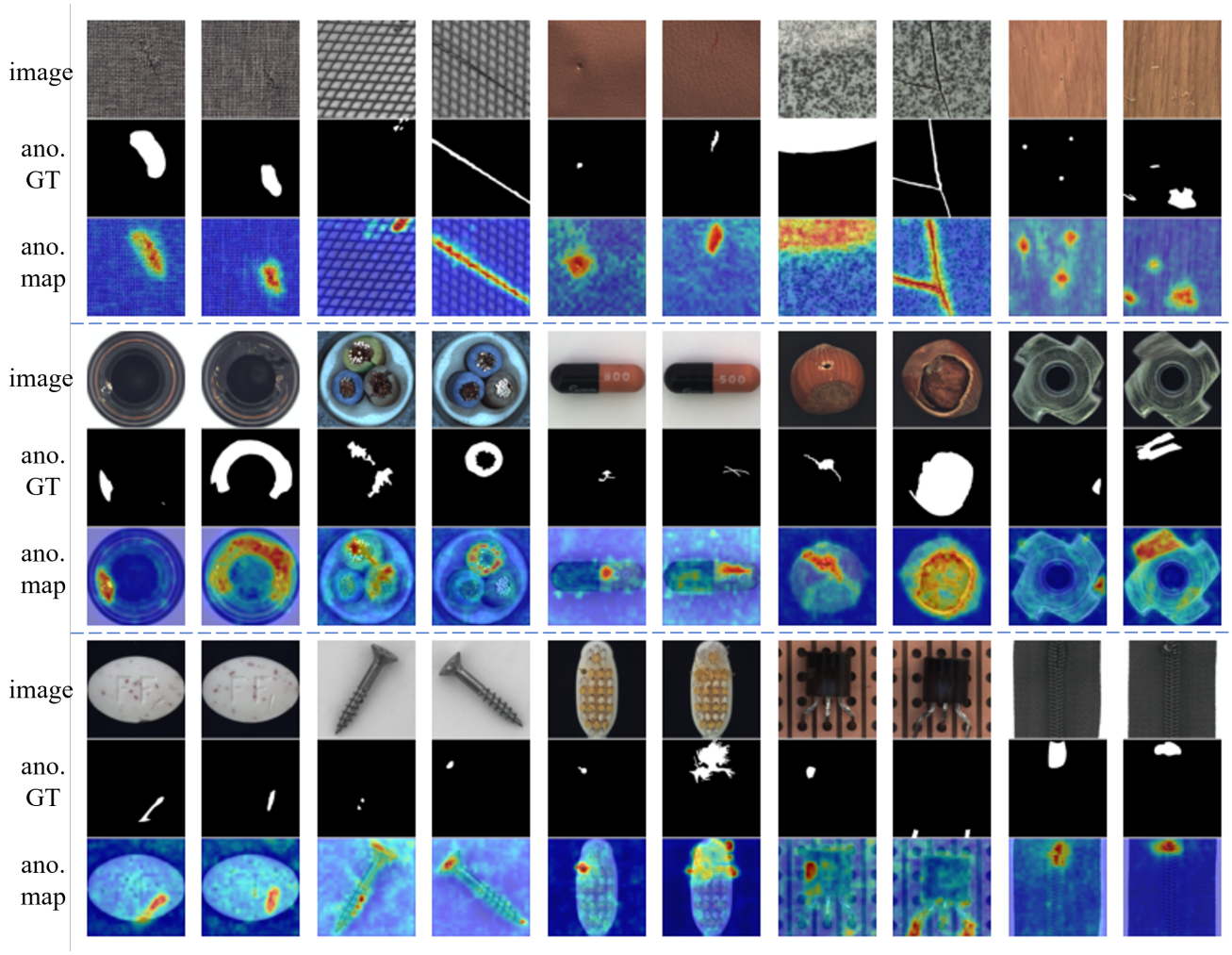


图 A1. MVTec-AD 上的异常区域可视化图。所有样本均为随机选取。

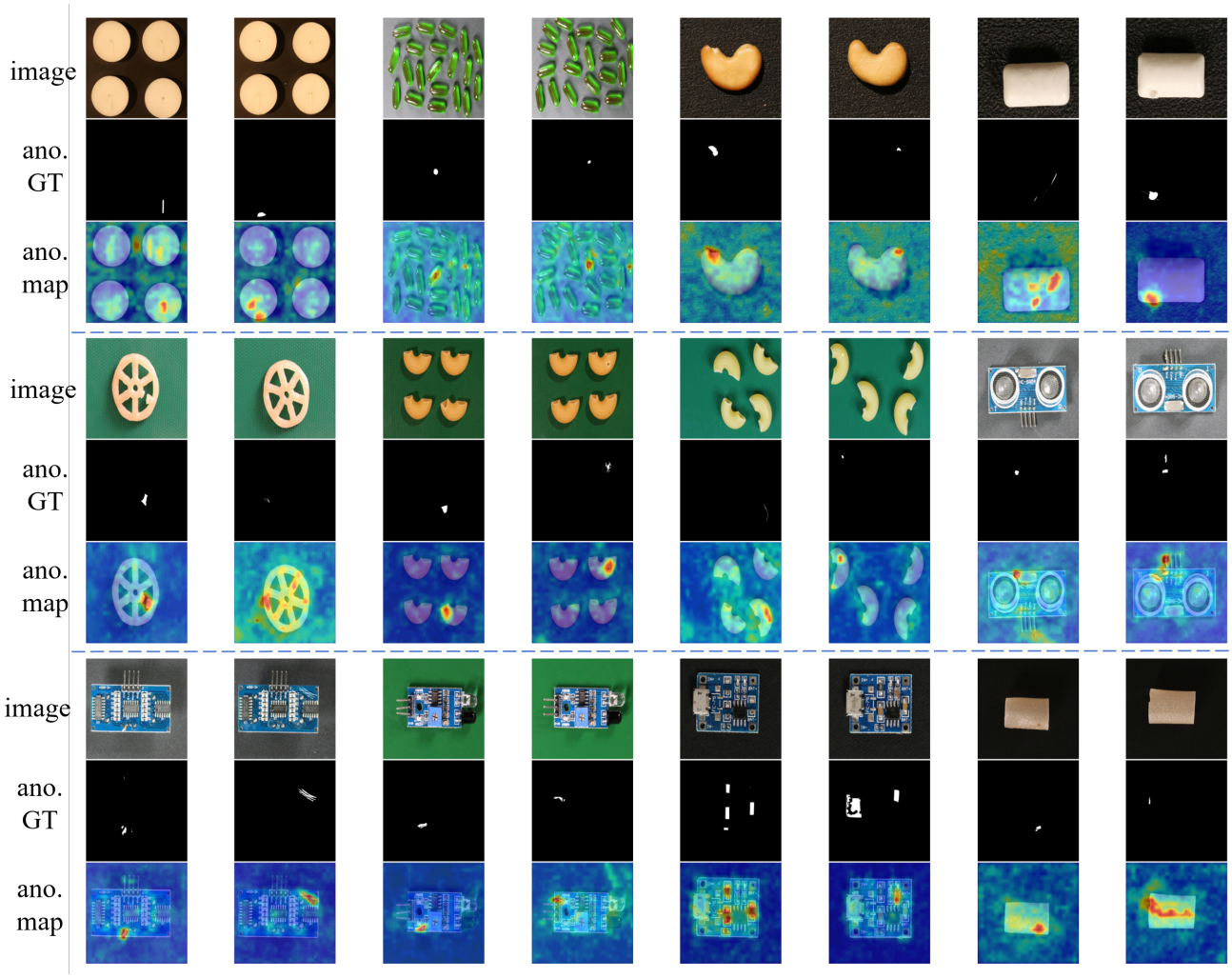
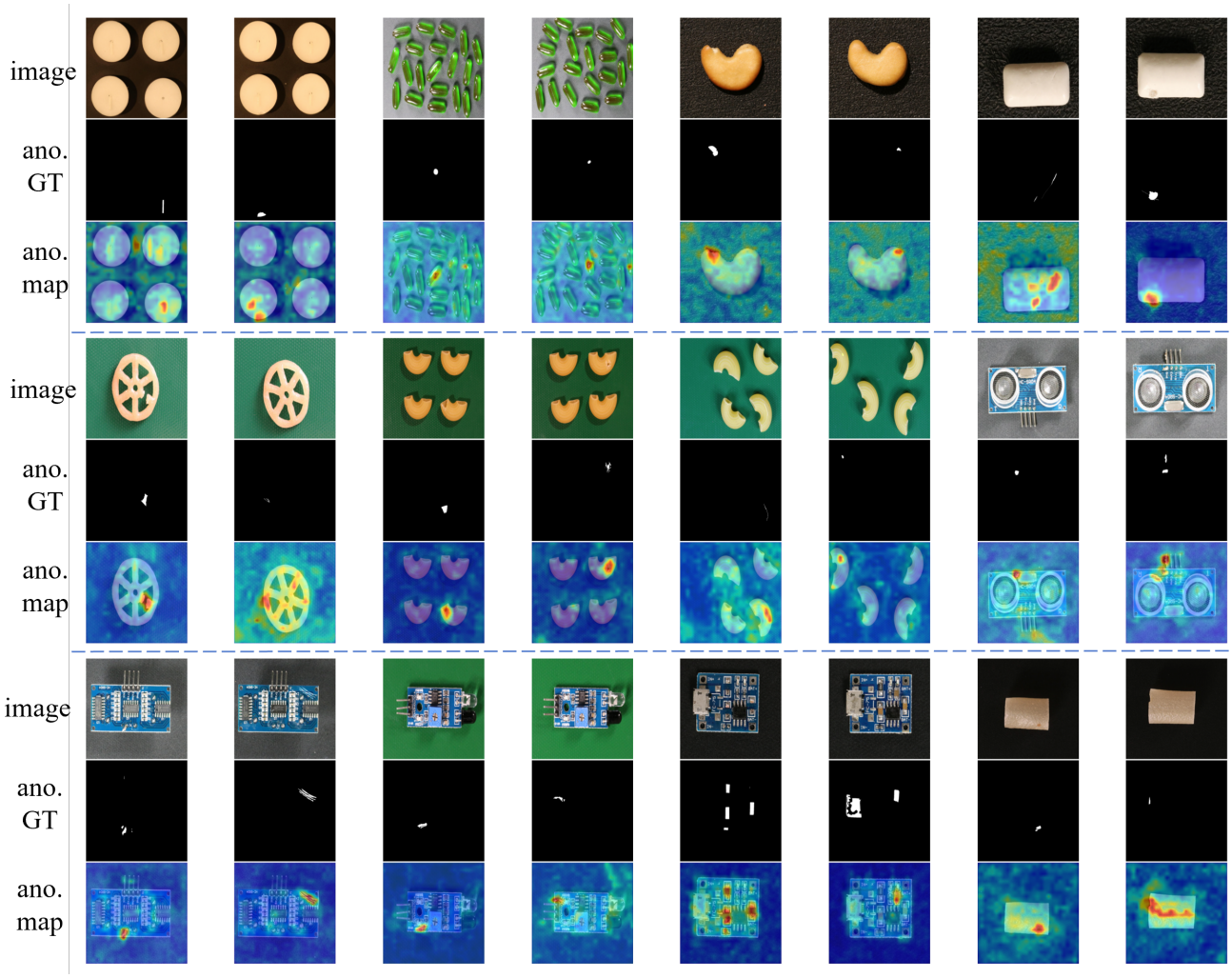


Figure A2. Anomaly maps visualization on VisA. All samples are randomly chosen.



图A2. VisA上的异常图可视化。所有样本均为随机选取。

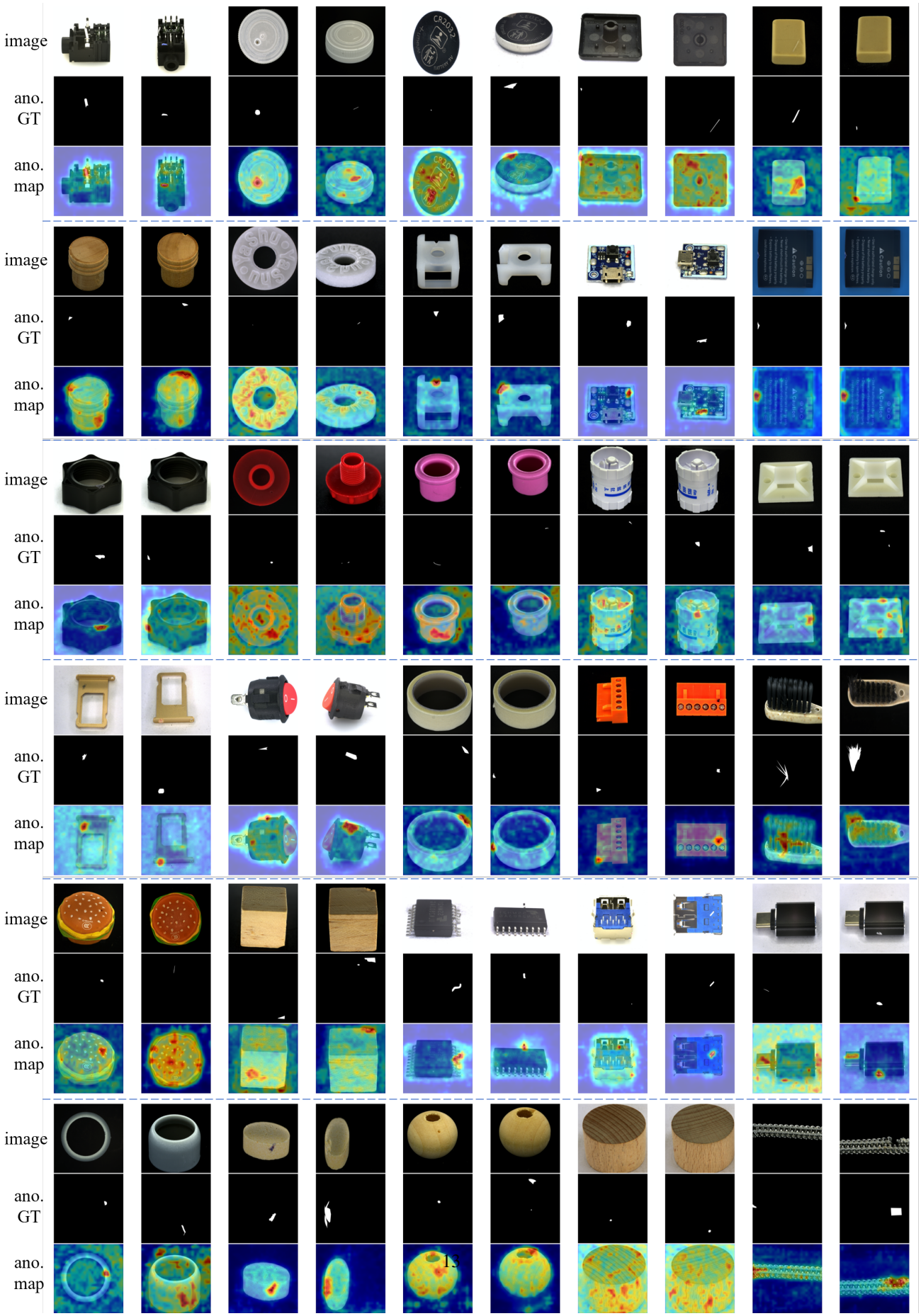
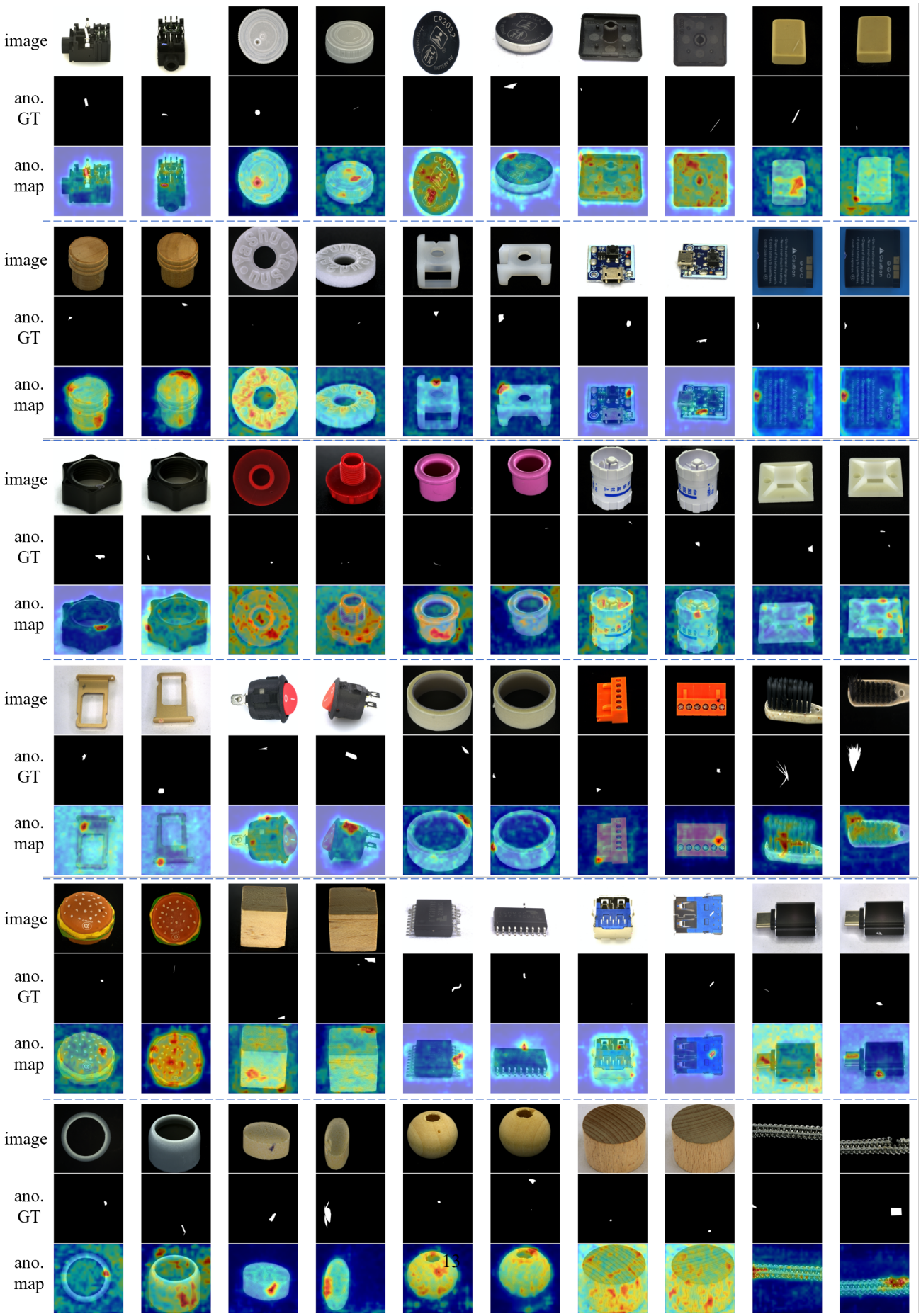


Figure A3. Anomaly maps visualization on Real-IAD. All samples are randomly chosen.



图A3. Real-IAD上的异常图可视化。所有样本均为随机选取。