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]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*

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-

✉ Corresponding author. † Former advisors when the project initiated.

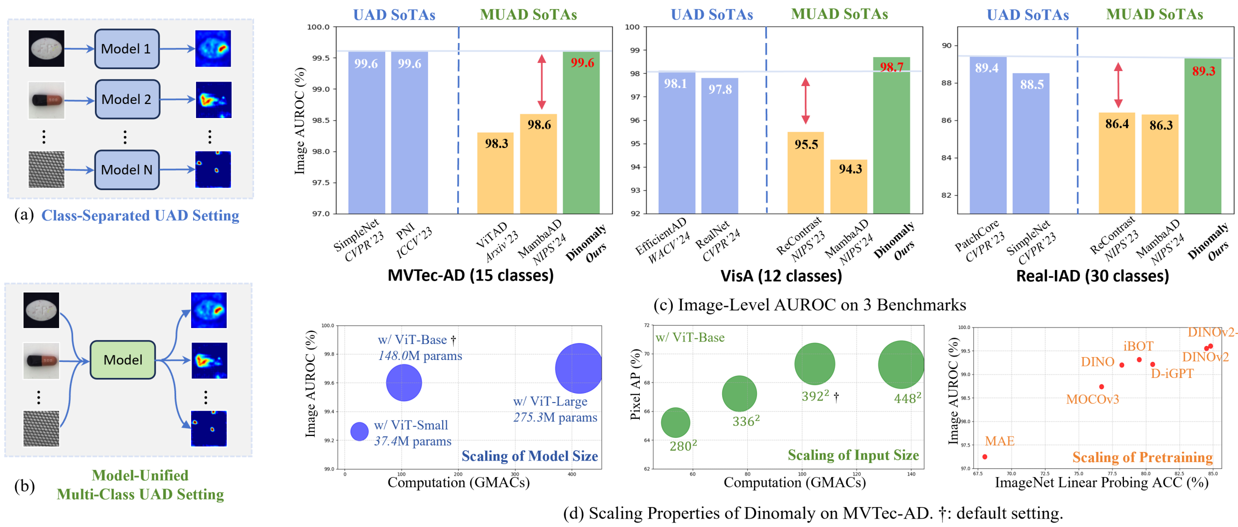


Figure 1. Setting, benchmarking, and scaling of Dinomaly. (a) Task setting of class-separated UAD. (b) Task setting of MUAD. (c) Comparison with previous SoTA methods on MVTec-AD [3], VisA [70], and Real-IAD [54]. (d) Scaling properties of Dinomaly.

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

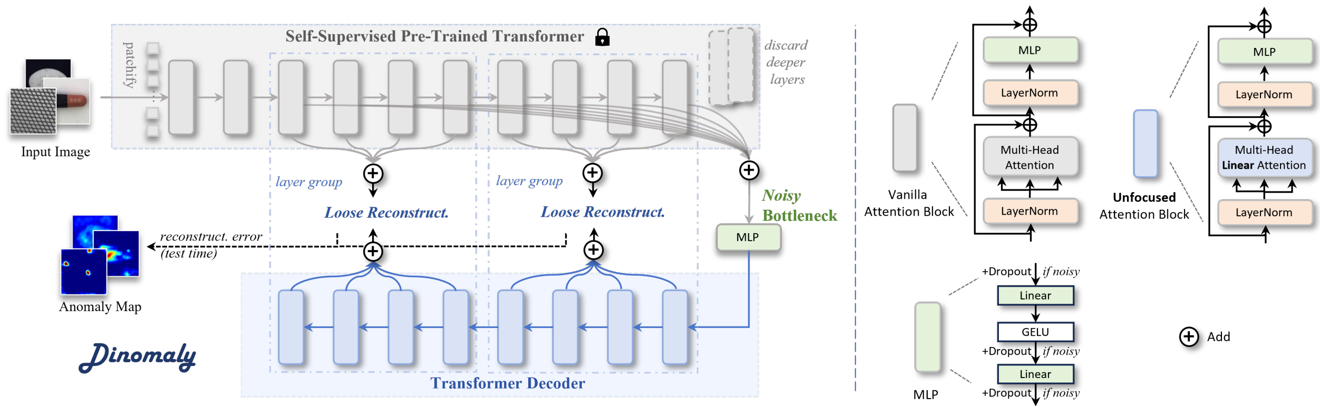


Figure 2. The framework of Dinomaly, built by simple and pure Transformer building blocks.

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.

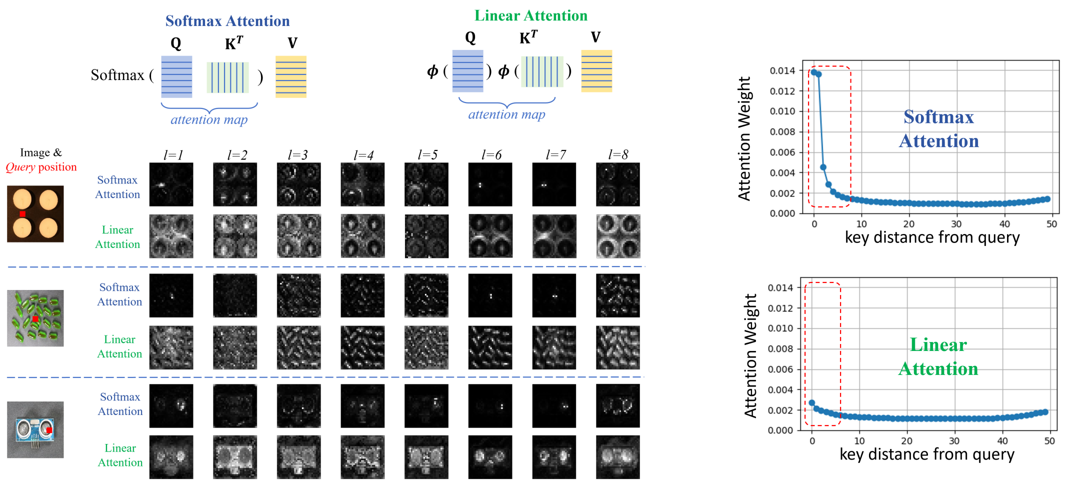
Foundation Transformers. Foundation models, especially ViTs [12, 33] pre-trained on large-scale datasets, serve as a basis and starting point for specific computer vision tasks. Such networks employ self-supervised learning schemes such as contrastive learning (MoCov3 [6], DINO [4]), masked image modeling (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-



(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 $\mathbf{V} \in \mathbb{R}^{N\times d}$:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^Q \, , \mathbf{K} = \mathbf{X}\mathbf{W}^K \, , \mathbf{V} = \mathbf{X}\mathbf{W}^V \, , \tag{1}$$

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]

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

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

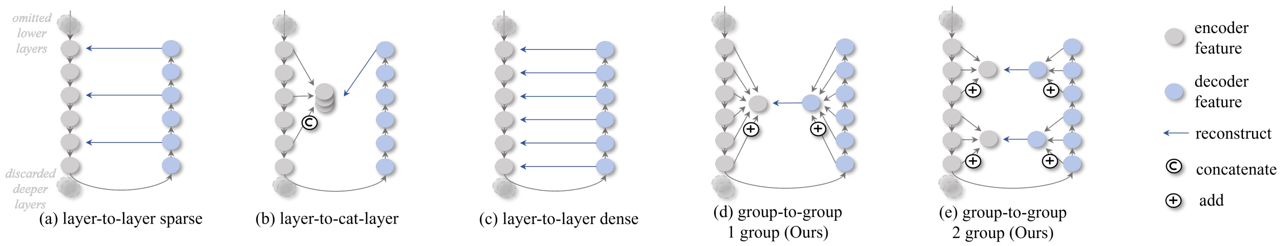


Figure 4. Schemes of reconstruction constraint. (a) Layer-to-layer (sparse). (b) Layer-to-cat-layer. (c) Layer-to-layer (dense). (d) Loose group-to-group, 1-group (Ours). (e) Loose group-to-group, 2-group (Ours).

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.

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

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-

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**

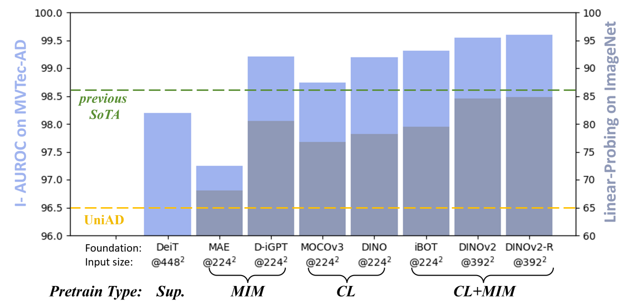


Figure 5. Image-level AUROC of Dinomaly equipped with various ViT foundations, and their linear-probing accuracy on ImageNet. MIM: Masked Image Modeling. CL: Contrastive Learning.

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.

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

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

[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. Draem-a 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 multiclass 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