HLRAD: High-dimensional Latent Representation for Unified Anomaly Detection

Tianrui Zhang, Ziheng Zang, and Ningmu Zou*

Nanjing University, Nanjing, China
{ztr,zangzh}@smail.nju.edu.cn, nzou@nju.edu.cn

Abstract. Unified unsupervised anomaly detection aims to train a single model to detect and localize diverse anomalies across multi-class samples, representing one of the most challenging tasks in anomaly detection. Existing methods typically rely on highly compressing inputs into a low-dimensional latent space to subsequently reconstruct anomalous features. However, low-dimensional latent spaces critically limit information capacity, leading to unavoidable feature loss during compression and significantly degrading the quality of reconstruction and representational power of the model. Therefore, we indicate that dimensionality compression in latent space is not a necessary requirement for anomaly detection, and the editability introduced by compression is not inherently essential for models reconstructing anomalies. Thus, we propose a novel anomaly detection framework without latent space compression, called **HLRAD** (**H**igh-dimensional **L**atent **R**epresentation for Unified **A**nomaly **D**etection). Unlike prior methods, HLRAD innovatively enables dimension expansion rather than compression in the latent space, and the method effectively avoids feature loss from compression to ensure high-quality reconstruction while constructing a semantically enriched high-dimensional latent representation space. Furthermore, by embedding the expanded high-dimensional semantic features into the latent representation, HLRAD enables the model to more fully capture the semantic feature distribution of normal samples, significantly enhancing complex anomaly detection performance. We conducted extensive experiments on major anomaly detection benchmark datasets, including MVTec-AD, VisA, and Real-IAD. In unified settings, HLRAD outperforms state-of-the-art methods.

Keywords: Unified Anomaly Detection · Representation Learning

1 Introduction

The aim of unsupervised anomaly detection is to learn the patterns of normal samples from a training set and identify anomaly samples that deviate from the normal distribution during inference [7, 9, 27]. It has been widely applied in industrial manufacturing [2, 15, 31], medical image analysis [1] and other domains [24, 27]. Conventional anomaly detection methods model each category

* Corresponding author.

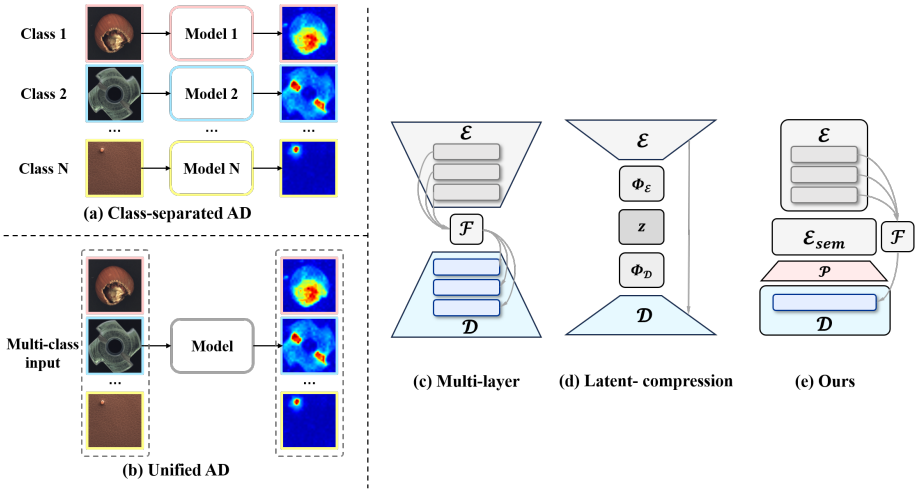

Fig. 1: Anomaly detection task settings and comparison of reconstruction paradigms. (a) Class-separated methods train separate models for each class. (b) Unified anomaly detection aims to handle all classes with a single model. (c) and (d) are two common paradigms: (c) multi-layer fused feature $\mathcal{F}$ matching performs equal-dimension layer-wise alignment between the encoder $\mathcal{E}$ and decoder $\mathcal{D}$, which is bounded by the encoder inherent dimension. (d) latent-compression methods compress features through the latent feature z into a low-dimensional space, inevitably incurring feature information loss. (e) Our method encodes features into a higher-dimensional space, enhancing reconstruction capacity without compression.

independently, requiring dedicated models for every object class. However, as the number of classes and anomaly types increases in real-world industrial scenarios, the per-class method incurs prohibitively high training and maintenance costs. Unified anomaly detection addresses this limitation by training a single model to handle multi-class simultaneously, offering a far more scalable and effective solution [27]. This paper therefore focuses on the promising yet challenging problem of multi-class unified anomaly detection.

As shown in Fig. 1, existing unified anomaly detection methods primarily rely on reconstruction [7, 9, 28], with models capable of reconstructing anomaly samples into normal to achieve anomaly detection during inference. However, two core challenges currently exist. The first is the shortcut learning problem [7, 27, 28]. The fundamental objective of unsupervised reconstruction training tends towards an identity mapping. Models may not effectively learn normal patterns but instead learn shortcuts to perfectly reconstruct inputs. This leads to anomaly samples being reconstructed via shortcuts during inference, thereby failing to be detected. UniAD [27] proposed a unified reconstruction framework based on Transformers [4], which mitigates shortcut learning tendencies by increasing reconstruction difficulty. Some methods further introduce feature depth damage, compelling the decoder to learn more robust representations of normal

patterns [7, 26]. While these methods alleviate the shortcut problem to some extent, generalised feature damage struggles to ensure the retention of crucial features.

Secondly, there is the issue of semantic fidelity in the representation space [8, 9]. Reconstruction-based anomaly detection requires models to perform semantically correct reconstructions of normal patterns, imposing stringent demands on the informational integrity of latent representations [8,9,28]. However, existing methods predominantly rely on continuous tokenizers [8,9] and discrete tokenizers [5, 16] to highly compress pixel space into latent space. Such highly compressed representations unavoidably discard substantial semantic and spatial detail, rendering it challenging for decoders to achieve semantically accurate reconstructions of normal patterns [30]. This severely compromises the accuracy of the detection. Methods operating in the raw dimensions of pre-trained features [3, 21] inherently constrain the decoder modeling capabilities to fixed feature dimensions. Regarding the above issues, we propose that the dimensionality compression commonly employed in existing methods is not an inherent requirement for anomaly detection tasks but rather constitutes a fundamental bottleneck constraining reconstruction quality. Based on this, we introduce a method contrary to the prevailing paradigm, which is to expand rather than compress the representation dimensions within the latent space. By extending the decoder working space, the model not only preserves the original feature information in its entirety, but more importantly, gains greater modeling capacity. This enables it to capture the complex distributions of normal patterns across different categories more accurately under a unified multi-class setting. To our knowledge, HLRAD provides an early exploration of dimension expansion and retaining semantic representations for anomaly detection.

Specifically, we propose a unified anomaly detection method called **HLRAD** (**H**igh-dimensional **L**atent **R**epresentation for unified **A**nomaly **D**etection). After the encoder processes the pixel-space features, we introduce a latent semantic branch that projects the encoder features into a higher-dimensional semantic space, extracting semantic representations through a semantic encoder. The embedded features are then projected into the high-dimensional decoding space via a dimension-expansion MLP with feature dropout, effectively blocking the direct mapping path from input to decoder, while simultaneously expanding the representation dimensionality to provide the decoder with sufficient modeling capacity. Within this high-dimensional latent space, a lightweight linear-attention decoder reconstructs the multi-layer feature representations of the encoder. Furthermore, we propose a Selective Gated Fusion mechanism that leverages the attention from the encoder to identify semantically salient tokens, and embeds the high-dimensional semantic features from the semantic branch into the reconstruction latent features through channel-wise adaptive gating, thereby enhancing normal semantic pattern modeling while preventing erroneous reconstruction of anomalous regions due to information leakage.

We conduct experiments on three widely used industrial anomaly detection benchmarks—MVTec-AD [2], VisA [31] and Real-IAD [24]. HLRAD surpasses

previous state-of-the-art unified frameworks under the unified multi-class setting. Our contributions can be summarized as follows:

1. We propose a novel dimensionality design for the latent space in the unified anomaly detection paradigm, and provide an exploration of a reconstruction framework based on dimensionality expansion.

2. We present the HLRAD unified anomaly detection framework, which extracts high-dimensional semantic representations through a latent semantic branch and integrates a dimension-expansion MLP with a selective gated fusion mechanism for semantic feature embedding, simultaneously expanding decoder modeling capacity and effectively blocking shortcut learning paths to achieve semantically faithful reconstruction of normal patterns.

3. Extensive experiments on benchmarks including MVTec-AD [2], VisA [31] and Real-IAD [24] demonstrate that HLRAD achieves state-of-the-art performance under the unified multi-class setting.

2 Related Work

Unified Anomaly Detection. The aim of unified anomaly detection is to build a single model capable of detecting anomalies across multiple classes, thereby avoiding the huge cost of training separate models for each class [7, 9, 27]. Under the unified setting, models are susceptible to the identity shortcut problem, where both normal and abnormal samples are faithfully reconstructed during inference [7, 14]. Moreover, models are prone to semantic confusion, in which the simultaneous training on diverse types of classes causes inputs to be reconstructed according to the semantic patterns of other classes [9]. UniAD [27] first introduces a Transformer-based reconstruction model that mitigates the identity shortcut by increasing the learning difficulty. Subsequently, HVQ-Trans [13] introduced a hierarchical vector quantisation framework to address the shortcut issue further. ReContrast [6] proposes a cross reconstruction method between dual encoders to mitigate category confusion. MambaAD [8] uses a state space model with various scanning strategies to extract local and global features.

Latent Compressed Anomaly Detection. Latent compressed representations encourage capturing structural patterns shared across categories, this paradigm has been widely adopted for unified multi-class anomaly detection [8, 9, 13, 26]. DiAD [9] is based on the Stable Diffusion VAE tokenizer [20] to compress pixel-space features into a low-dimensional latent space to make semantic features guide the diffusion decoder. HVQ-Trans [13] uses a hierarchical VQ codebook [16] as a tokenizer to encode pixel space images into multi-level discrete codebook indexes. PMAD [26] uses a discrete VAE [10] to compress images into discrete visual tokens to reconstruct the masked tokens. While low-dimensional latent spaces endow generative models with enhanced editability and flexibility, these methods result in significant feature loss during compression. Our method avoids the need for latent space compression entirely, thus overcoming this limitation and preserving richer feature information.

High-dimensional Representations for Anomaly Detection. Several recent unified anomaly detection methods avoid explicit latent space compression by directly using pretrained models [7, 14, 28]. ViTAD [28] employs a frozen ViT encoder [4] paired with a trainable decoder to reconstruct intermediate features. Dinomaly [7] further introduces a bottleneck MLP with dropout, to compel the decoder to learn more robust normal-pattern representations. While operating in the original feature dimensionality effectively preserves spatial details, it also constrains model capacity in the fixed pretrained dimension. Recent advances in image generation have revealed the advantages of high-dimensional representation spaces. RAE [30] proposes that high-dimensional latent spaces yield semantically richer and more coherent representations than low-dimensional VAE counterparts. DDT [25] proposes that a shallow and wide decoder can efficiently handle such high-dimensional features. Building on these insights [25, 30], we propose an anomaly detection framework that expands the width and depth of the decoder, enhancing its ability to model normal features and enabling more accurate anomaly reconstruction without dimensionality compression.

3 Method

3.1 Overview

In this paper, we propose a unified multi-class anomaly detection framework based on multi-scale feature reconstruction. To enrich the latent feature representation, we introduce a frozen semantic Transformer encoder to re-encode patch tokens into a higher-dimensional semantic space, while preserving fine-grained information via a selective gating mechanism. To reconstruct multi-scale feature representation, we further introduce a high-capacity decoder whose hidden dimension exceeds that of the teacher encoder. As illustrated in Fig. 2, the proposed framework consists of four stages: (i) multi-scale feature extraction with a semantic encoding branch, (ii) selective gated fusion of texture and semantic features, (iii) high-dimensional decoding through a projection MLP, and (iv) cosine-distance-based anomaly scoring in inference.

3.2 Semantic Feature Extraction

Existing Reconstruction-based anomaly detection methods [3,7,27] typically employ the raw features of the frozen encoder directly as the input and reconstruction target of the decoder. However, these approaches are constrained by the encoder inherent feature dimension d_t, feeding features into the decoder through either information compression or dimension-preserving mappings, which limits the expressiveness of the feature space for complex semantic structures. Moreover, pre-trained encoders [17] are trained on natural image datasets, whereas anomaly detection tasks involve industrial features and objects that deviate significantly from the pre-training distribution. This distribution shift causes the extracted latent features to undergo systematic drift, thereby degrading downstream reconstruction quality.

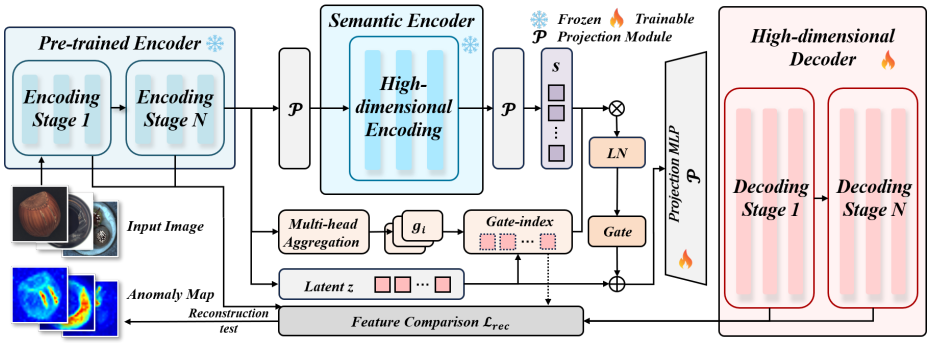

Fig. 2: Overall framework of the proposed method. The input image is processed by encoder for multi-stage feature extraction and a semantic encoder for high-dimensional encoding. The selective gated fusion adaptively combines semantic features. The fused features are projected into a high-dimensional latent space via a dimension-expanding MLP and reconstructed by a trainable high-dimensional decoder.

To address these issues, we propose leveraging a frozen Diffusion Transformer [18, 30] encoder to transform features into a higher-dimensional semantic space, where they are re-encoded to yield richer semantic representations. Specifically, as illustrated in Fig. 2, we adopt the frozen DiT encoder $\mathcal{D}_{\text{enc}}$ of depth D_e from a pre-trained RAE [30]. The fused patch tokens $\bar{\mathbf{F}} \in \mathbb{R}^{N \times d_t}$ extracted by the encoder are first projected into the DiT high-dimensional latent space via a learnable linear mapping:

$$\mathbf{S}_e^{(0)} = \phi_{\text{in}}(\bar{\mathbf{F}}) + \mathbf{E}_{\text{pos}} \in \mathbb{R}^{N \times d_s}, \tag{1}$$

where ϕ_{in} is a mapping module that projects from $\mathbb{R}^{d_t}$ to $\mathbb{R}^{d_s}$, d_s denotes the DiT hidden dimension which exceeds patch token dimension d_t, and $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{N \times d_s}$ denotes sinusoidal positional embeddings. The projected sequence is then processed through D_e depth blocks and can be written as:

$$\mathbf{S}_e^{(j)} = \mathcal{B}_j\left(\mathbf{S}^{(j-1)},\ \mathbf{c}\right), \quad j \in \{1, 2, \ldots, D_e\}, \tag{2}$$

where $\mathbf{c} \in \mathbb{R}^{d_s}$ is an unconditional learnable embedding, and each block $\mathcal{B}_j$ incorporates Rotary Position Embedding [23]. The final DiT output is projected back to the original dimension:

$$\mathbf{S} = \phi_{\text{out}}\left(\mathbf{S}^{(D_e)}\right) \in \mathbb{R}^{N \times d_t}, \tag{3}$$

where ϕ_{out} reduces the dimensionality from $\mathbb{R}^{d_s}$ to $\mathbb{R}^{d_t}$. All parameters of $\{\mathcal{B}_j\}_{j=1}^{D_e}$ remain frozen, only ϕ_{in}, ϕ_{out}, and $\mathbf{c}$ are learnable. The DiT encoder operates at dimension $d_s > d_t$, providing richer representational capacity to encode hierarchical semantic structures that the original d_t dimensional space cannot fully express. Since the DiT is pre-trained on large-scale image generation, its intermediate representations carry complementary semantic priors that facilitate

capturing the global semantic coherence required for detecting structural anomalies across diverse object categories. In addition, the DiT encoder was originally trained as a denoising backbone and thus possesses inherent robustness to distributional perturbations. It treats the feature drift induced by out-of-distribution data as a form of input noise, that mitigate the adverse effects of domain gap on reconstruction quality without requiring any domain-specific fine-tuning.

3.3 Selective Gated Fusion

The semantic features $\mathbf{S} \in \mathbb{R}^{N \times d_t}$ extracted by the DiT encoder and the original features $\bar{\mathbf{F}} \in \mathbb{R}^{N \times d_t}$ share the same dimensionality but encode fundamentally different information. The former carries global structural and semantic priors after re-encoding through a high-dimensional semantic space, while the latter preserves the fine-grained textural and local spatial information native to the encoder. Uniformly replacing the original features with semantic features at all spatial positions would sacrifice critical textural cues for detecting subtle local anomalies, whereas entirely discarding the semantic features would forgo the global semantic understanding that the original features lack. Thus, a mechanism is needed to selectively inject semantic information at appropriate spatial locations while retaining the original textural representations elsewhere.

Therefore, we propose a Selective Gated Fusion (SGF) module that leverages the encoder attention to identify semantically salient regions and injects semantic features only at those positions. Specifically, we extract the multi-head self-attention weights $\mathbf{A} \in \mathbb{R}^{N_h \times N}$ from the last encoder target layer and obtain per-patch semantic importance scores via head-wise averaging:

$$\alpha_i = \frac{1}{N_h} \sum_{h=1}^{N_h} \mathbf{A}^{(h)}[i], \quad i \in \{1, 2, \ldots, N\}, \tag{4}$$

where N_h is the number of attention heads in layer h. These attention weights naturally encode each patch contribution to the global image representation, with high-attention regions corresponding to semantically salient locations such as boundaries and structural keypoints.

During training, we uniformly sample a gating ratio r from $[0, r_{\max}]$ and select the k positions with the highest importance scores, where $k = \lfloor r \cdot N \rfloor$, to form the gating set $\mathcal{K}$. For each selected position $i \in \mathcal{K}$, a channel-wise fusion weight is computed by concatenating the features from both branches and passing them through a linear layer characterized by sigmoids σ:

$$\mathbf{g}_i = \sigma\big(\mathbf{W}_g \left[\bar{\mathbf{F}}_i \,\|\, \mathbf{S}_i\right] + \mathbf{b}_g\big) \in [0,1]^{d_t}, \tag{5}$$

where $[\cdot\|\cdot]$ denotes channel-wise concatenation and $\mathbf{W}_g \in \mathbb{R}^{d_t \times 2d_t}$, $\mathbf{b}_g \in \mathbb{R}^{d_t}$ are learnable parameters. The fused features are then given by:

$$\hat{\mathbf{F}}_i = \begin{cases} (1 - \mathbf{g}_i) \odot \bar{\mathbf{F}}_i \;+\; \mathbf{g}_i \odot \mathbf{S}_i, & i \in \mathcal{K}, \\ \bar{\mathbf{F}}_i, & i \notin \mathcal{K}. \end{cases} \tag{6}$$

This mechanism enables spatial positions with strong semantic demands to fully incorporate high-dimensional semantic features through gating, while texture-dominant positions retain the original features without interference. Random gating during training thus serves as a form of structured regularization that prevents the decoder from becoming overly reliant on semantic features while exposing it to semantically enriched representations during a random subset of training steps.

3.4 High-dimensional Decoder

Existing reconstruction-based anomaly detection methods [3,7,27] typically employ decoders with the same dimensionality as the encoder, where the bottleneck layer serves solely for information compression, and the limited dimensionality constrains the decoder capacity for modeling complex non-linear mappings. DDT [25] and RAE [30] have shown that a shallow and wide decoder operating in a high-dimensional latent space is more efficient than a deep and narrow architecture. We find that this finding aligns naturally with the framework requirements of anomaly detection. In our framework, the latent features $\bar{\mathbf{F}} \in \mathbb{R}^{N \times d_t}$ and the semantic features $\mathbf{S}_e \in \mathbb{R}^{N \times d_s}$ originate from encoding processes in two distinct dimensional spaces, and the decoder must accommodate and reconstruct both types of information within a unified high-dimensional space. Consequently, we adopt the DDT decoupled decoder design [25], setting the hidden dimension d_d to substantially exceed d_t while moderately increasing the depth of the decoder relative to the original DDT decoder to improve the reconstruction capacity for the anomaly detection task.

Dimension-expanding MLP. Before entering the decoder, fused features $\hat{\mathbf{F}} \in \mathbb{R}^{N \times d_t}$ are mapped into the high-dimensional space of the decoder through a Dimension-expanding MLP with triple dropout regularization:

$$\mathbf{H} = \mathrm{Drop}_p\Big(\mathbf{W}_2\ \mathrm{GELU}\Big(\mathrm{Drop}_p\Big(\mathbf{W}_1\ \mathrm{Drop}_p\Big(\hat{\mathbf{F}}\Big)\Big)\Big)\Big) \in \mathbb{R}^{N \times d_d}, \tag{7}$$

where $\mathbf{W}_1 \in \mathbb{R}^{4d_d \times d_t}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_d \times 4d_d}$ are learnable weights, and $\mathrm{Drop}_p(\cdot)$ denotes dropout with probability p applied in pre-projection, post-activation, and post-projection. Unlike conventional bottlenecks that merely compress information, this design expands features from d_t to d_d dimensions, providing the decoder with a high-dimensional working space while introducing stochastic perturbations that prevent identity mapping. The triple dropout forces the decoder to learn generalizable normal-pattern reconstruction rather than memorizing training-specific features, ensuring that anomalous inputs unseen during training produce significant reconstruction errors.

Decoder Training and Loss. The bottleneck output is subsequently processed through D_d standard blocks operating at dimension d_d. To further prevent co-adaptation among attention heads, we randomly drop a fraction p_h of attention heads during each forward pass at training time. Specifically, a Bernoulli mask $\mathbf{m} \in \{0,1\}^{N_h}$ is sampled with probability $1 - p_h$, and the remaining heads

are rescaled by $N_h / \sum_j \mathbf{m}_j$ to preserve the expected output norm. This mechanism complements the dropout along a different dimension from the Dimension-expanding MLP, jointly enhancing the decoder ability to generalize beyond the seen normal patterns. Following the reverse distillation paradigm [7], the multi-layer decoder features are grouped and projected back to the teacher dimension d_t via per-group linear projections with LayerNorm. The model is trained with a global cosine similarity loss between the projected decoder features and the corresponding frozen teacher features, with stop-gradient applied on the teacher side. During inference, the per-position cosine distance between teacher and decoder features serves directly as the pixel-level anomaly map, which is upsampled to the input resolution.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate our method on three common anomaly detection benchmarks: MVTec-AD [2], VisA [31] and Real-IAD [24]. MVTec-AD [2] comprises 15 categories (5 texture and 10 object classes), with 3,629 normal images for training and 1,725 test images (467 normal, 1,258 anomalous). VisA [31] covers 12 object categories. We follow the official split, 8,659 normal images form the training set and 2,162 images constitute the test set (962 normal, 1,200 anomalous). Real-IAD [24] is a recently released large-scale benchmark spanning 30 distinct objects. Using the official all-view split, it yields 36,465 normal training images and 114,585 test images (63,256 normal, 51,329 anomalous).

Metrics. We employ the Area Under the Receiver Operating Characteristic Curve (AUROC), Average Precision (AP), and F1-score maximum ($F1_{\max}$) to evaluate both anomaly detection and localization performance.

Implementation Details. We adopt DINOv3-ViT-L/16 [22] (d_t=1024, 24 layers) as the frozen encoder, and a 28-block DiT encoder (d_s=1152) pre-trained from RAE [30] as the semantic encoder. The decoder consists of 8 Transformer blocks with hidden dimension d_d=2048 and 32 attention heads. Eight evenly spaced layers $\{4, 6, 8, 10, 12, 14, 16, 18\}$ are selected as target layers from the DINOv3 encoder, and features are fused into 2 groups for both the encoder and decoder sides. The input image is first resized to 512^2 and then center-cropped to 448^2, yielding a 28×28 patch sequence. The bottleneck dropout rate is set to 0.3 by default and increases to 0.4 on the more diverse Real-IAD. The maximum gating ratio r of the SGF module is 0.5. The DropHead rate p_h is 0.1. We use AdamW [12] with AMSGrad, weight decay 10^{-4} and a cosine learning rate schedule decaying to 2×10^{-4} after a 100-iteration linear warmup. Gradient norms are clipped to 0.1. The network is trained for 20k iterations, using 8×NVIDIA A100 GPUs.

4.2 Main Results

Quantitative Results. We compare our method with several state-of-the-art (SOTA) multi-class anomaly detection approaches, including reconstruction-

Table 1: Comparison with multi-class unsupervised anomaly detection methods on MVTec-AD [2], VisA [31], and Real-IAD [24]. The best performance is in bold, and the second-best is underlined.

Method	RD4AD [3]	UniAD [27]	SimpleNet [11]	DeSTSeg [29]	DiAD [9]	MambaAD [8]	ViTAD [28]	Dinomaly [7]	INP-Former [14]	**Ours**
Venue	*CVPR'22*	*NIPS'22*	*CVPR'23*	*CVPR'23*	*AAAI'24*	*NIPS'24*	*CVIU'25*	*CVPR'25*	*CVPR'25*	
MVTec-AD										
I-AUROC	94.6	96.5	95.3	89.2	97.2	98.6	98.3	<u>99.6</u>	**99.7**	**99.7**
I-AP	96.5	98.8	98.4	95.5	99.0	99.6	99.4	<u>99.8</u>	**99.9**	**99.9**
I-F1$_{max}$	95.2	96.2	95.8	91.6	96.5	97.8	97.3	99.0	<u>99.2</u>	**99.3**
P-AUROC	96.1	96.8	96.9	93.1	96.8	97.7	97.7	98.4	<u>98.5</u>	**98.6**
P-AP	48.6	43.4	45.9	54.3	52.6	56.3	55.3	69.3	<u>71.0</u>	**71.3**
P-F1$_{max}$	53.8	49.5	49.7	50.9	55.5	59.2	58.7	69.2	<u>69.7</u>	**70.7**
VisA										
I-AUROC	92.4	88.8	87.2	88.9	86.8	94.3	90.5	98.7	<u>98.9</u>	**99.0**
I-AP	92.4	90.8	87.0	89.0	88.3	94.5	91.7	98.9	**99.0**	**99.0**
I-F1$_{max}$	89.6	85.8	81.8	85.2	85.1	89.4	86.3	96.2	**96.6**	**96.6**
P-AUROC	98.1	98.3	96.8	96.1	96.0	98.5	98.2	98.7	<u>98.9</u>	**99.1**
P-AP	38.0	33.7	34.7	39.6	26.1	39.4	36.6	<u>53.2</u>	51.2	**55.5**
P-F1$_{max}$	42.6	39.0	37.8	43.4	33.0	44.0	41.1	<u>55.7</u>	54.7	**57.5**
Real-IAD										
I-AUROC	82.4	83.0	57.2	82.3	75.6	86.3	82.7	89.3	**90.5**	<u>90.1</u>
I-AP	79.0	80.9	53.4	79.2	66.4	84.6	80.2	86.8	**88.1**	<u>87.5</u>
I-F1$_{max}$	73.9	74.3	61.5	73.2	69.9	77.0	73.7	80.2	**81.5**	<u>81.1</u>
P-AUROC	97.3	97.3	75.7	94.6	88.0	98.5	97.2	<u>98.8</u>	**99.0**	**99.0**
P-AP	25.0	21.1	2.8	37.9	2.9	33.0	24.3	42.8	<u>47.5</u>	**47.7**
P-F1$_{max}$	32.7	29.2	6.5	41.7	7.1	38.7	32.3	47.1	<u>50.3</u>	**50.5**

based methods RD4AD [3], UniAD [27], DiAD [9], ViTAD [28], MambaAD [8], Dinomaly [7], and INP-Former [14], as well as embedding-based methods SimpleNet [11] and DeSTSeg [29]. The quantitative experimental results on the 3 datasets are presented in Tab. 1. On MVTec-AD, our method achieves the best performance across all metrics, with image-level scores of 99.7/99.9/99.3 and pixel-level scores of 98.6/71.3/70.7. Compared to the second-best results, the pixel-level AP and F1$_{\max}$ are improved by 0.3↑ and 1.0↑, respectively, demonstrating a clear advantage in anomaly localization. On VisA, our method reaches 99.0/99.0/96.6 at the image level and 99.1/55.5/57.5 at the pixel level. This corresponds to improvements of 0.1↑ on image-level AUROC and 0.2↑/2.3↑/1.8↑ at the pixel level over the second-best results. On Real-IAD, our method ranks second on image-level metrics with 90.1/87.5/81.1, and achieves the best and tied-best pixel-level performance of 99.0/47.7/50.5. Across the three datasets, our method achieves the best overall performance on most metrics, especially on pixel-level anomaly localization. The consistently strong performance across all three datasets validates the effectiveness and robustness of the proposed approach.

Qualitative Results. To further demonstrate the qualitative results, as shown in Fig. 3, we provide qualitative visualization of detection results, and compare against DiAD [9]. It can be intuitively observed that DiAD exhibits notable errors in certain complex anomalous regions, whereas our method accurately computes anomaly scores. This indicates that DiAD does not fully learn the normal patterns of these categories, and instead reconstructs certain normal

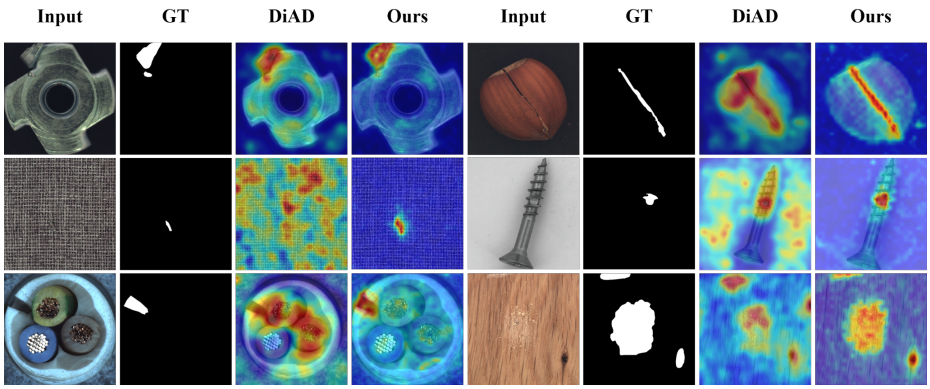

Fig. 3: Qualitative results comparison. Each group shows the input image, ground-truth mask (GT), and anomaly score maps produced by DiAD [9] and our method. Our method generates more focused responses that closely match the ground-truth anomaly boundaries.

region features as anomaly. The qualitative results intuitively confirm that our method effectively captures normal patterns and locates anomalous regions.

4.3 Tokenizer Representation Study.

To evaluate whether the dimensional-expanding tokenizer better preserves information, we compare it with the DiAD [9] VAE tokenizer on the MVTec-AD [2] training set. Both tokenizers are tested using only their encoder and decoder components under the same spatial resolution. Our tokenizer follows the RAE setting [30], using a frozen DINOv3 encoder [22] and a learned ViT decoder [4], while DiAD uses its officially released fine-tuned autoencoder. We report the mean per-category L1 loss, L2 loss, and cosine distance over all normal training images from the 15 MVTec-AD categories. As shown in Fig. 4 and Tab. 2, our tokenizer achieves lower reconstruction error across nearly all categories. On average, it reduces the L1 loss by 66.4% (0.0213 *vs.* 0.0633), the L2 loss by 90.0% (0.0010 *vs.* 0.0100), and the cosine distance by 91.8% (0.0023 *vs.* 0.0280), indicating better preservation of semantic structure.

4.4 Ablation Study

Overall Ablation. We conduct experiments to verify the effectiveness of the four proposed components, namely the Semantic Encoder (S_{enc}), Selective Gated Fusion (SGF), Dimensional-expanding MLP (DM) and High-dimensional Decoder (HD). DM maps features from the encoder dimension d_t to the decoder dimension d_d via a two-layer MLP with triple dropout regularization, serving as both a dimensional bridge and a stochastic bottleneck that prevents identity

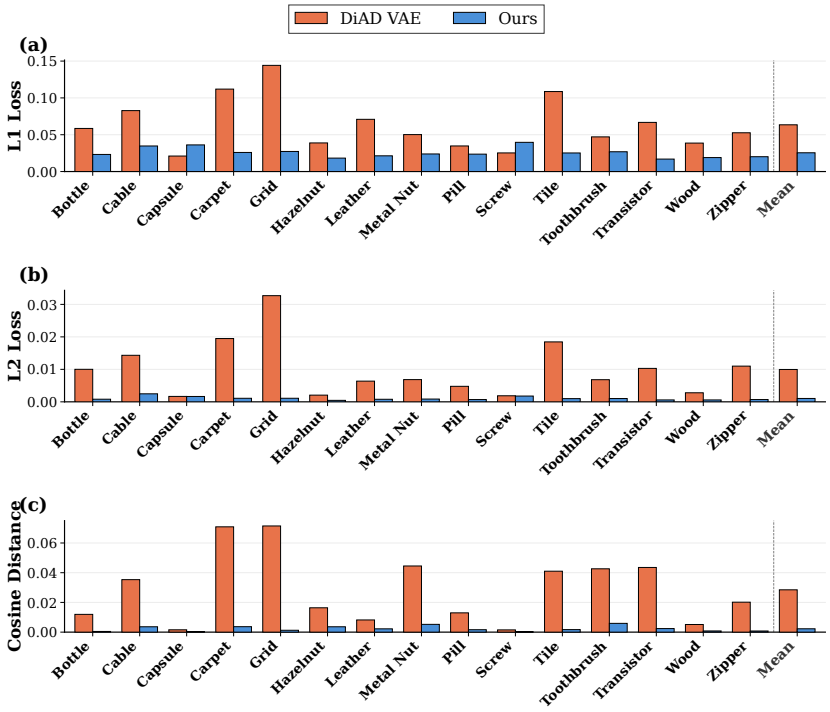

Fig. 4: Loss results comparison. Per-category reconstruction loss comparison between our dimensional-expanding tokenizer and the DiAD VAE [9] tokenizer on MVTec-AD [2], evaluated under the same spatial resolution. (a) L1 loss, (b) L2 loss, and (c) Cosine distance. Our tokenizer consistently achieves lower reconstruction error.

mapping. The baseline employs a reconstruction framework without augmentation of semantic features. The results are shown in Tab. 3. The Semantic Encoder S_{enc} provides a notable improvement over the baseline, validating the effectiveness of high-dimensional semantic encoding. Incorporating DM on top of S_{enc} yields a further improvement, as the regularization in the dimensional-expanding MLP prevents the decoder from learning identity shortcuts. The addition of HD further amplifies the gain, confirming that semantically enriched features are more effectively exploited within a high-dimensional reconstruction space. SGF complements these components by injecting semantic information at salient spatial positions, and the full combination of all four components achieves the best overall performance across both datasets.

Model Depth and Dimension. We study the decoder dimension d_d and depth D_d in Tab. 4. With D_d=8, increasing d_d generally improves performance, especially on pixel-level AP and F1, showing that a wider latent reconstruction space benefits anomaly localization. When fixing d_d=2048, increasing the depth from 4

Table 2: Mean reconstruction loss of different tokenizers on MVTec-AD. The best performance is in bold.

Loss	DiAD VAE [9]	Ours	Reduction
L1 Loss	0.0633	**0.0213**	66.4%
L2 Loss	0.0100	**0.0010**	90.0%
Cosine Distance	0.0280	**0.0023**	91.8%

Table 3: Overall ablation study on MVTec-AD and VisA. S_{enc} denotes the Semantic Encoder, SGF denotes Selective Gated Fusion, DM denotes the Dimensional-expanding MLP, and HD denotes the High-dimensional Decoder. The best performance is in bold.

S_{enc}	SGF	DM	HD	MVTec-AD						VisA					
				I-AUROC	I-AP	I-F1$_{\text{max}}$	P-AUROC	P-AP	P-F1$_{\text{max}}$	I-AUROC	I-AP	I-F1$_{\text{max}}$	P-AUROC	P-AP	P-F1$_{\text{max}}$
				99.6	99.8	99.0	98.4	69.3	69.2	98.7	98.9	96.2	98.7	53.2	55.7
✓				99.6	99.7	98.5	98.4	70.4	69.9	98.9	99.0	96.4	98.8	54.0	56.5
✓		✓		99.7	99.8	98.7	98.5	71.0	70.2	98.9	99.1	96.6	98.9	54.7	56.7
✓		✓	✓	99.7	99.9	99.2	98.5	71.2	70.5	99.1	99.2	96.9	99.0	55.3	57.2
✓	✓	✓		99.7	99.9	99.1	98.5	71.0	70.3	99.1	99.2	96.8	98.9	55.1	57.1
✓	✓	✓	✓	**99.7**	**99.9**	**99.3**	**98.6**	**71.3**	**70.7**	**99.0**	**99.0**	**96.6**	**99.1**	**55.5**	**57.5**

Table 4: Model scalability study on MVTec-AD and VisA. The upper portion varies the decoder dimension with fixed depth D_d=8. The lower portion varies the decoder depth with fixed dimension d_d=2048. The best performance is in bold.

d_d	D_d	MVTec-AD						VisA					
		I-AUROC	I-AP	I-F1$_{\text{max}}$	P-AUROC	P-AP	P-F1$_{\text{max}}$	I-AUROC	I-AP	I-F1$_{\text{max}}$	P-AUROC	P-AP	P-F1$_{\text{max}}$
768	8	99.5	99.7	98.4	98.3	70.0	69.5	98.8	98.9	96.0	98.7	53.5	56.1
1024	8	99.6	99.8	98.8	98.4	70.8	69.9	**99.0**	98.9	96.5	98.9	55.3	57.3
1536	8	99.6	99.8	99.0	98.5	71.1	70.3	**99.0**	98.9	96.5	**99.1**	**55.5**	57.3
2048	8	**99.7**	**99.9**	**99.3**	**98.6**	**71.3**	**70.7**	**99.0**	**99.0**	**96.6**	**99.1**	**55.5**	**57.5**
2048	4	99.6	99.8	98.8	98.4	71.0	70.0	98.8	98.9	96.5	98.8	54.5	57.0
2048	6	99.6	99.8	99.0	98.6	71.2	70.3	98.9	98.9	96.5	99.0	54.9	57.1
2048	8	**99.7**	**99.9**	**99.3**	98.6	**71.3**	**70.7**	**99.0**	**99.0**	**96.6**	99.1	55.5	**57.5**
2048	10	99.6	99.8	99.1	**98.7**	**71.3**	**70.7**	98.9	98.9	96.9	**99.1**	**55.6**	**57.5**

to 8 improves most metrics, while further increasing brings only marginal gains. Therefore, we use d_d=2048 and D_d=8 as the default setting, which provides the best overall trade-off between reconstruction capacity and training stability.

Semantic Encoder. We evaluate the semantic encoder S_{enc} settings, altering its model and training approach, which in turn affects its depth and dimensionality. As shown in Tab. 5, firstly, freezing consistently outperforms fine-tuning for both architectures. A trainable semantic encoder tends to co-adapt with the decoder, reducing its ability to provide complementary semantic information and weakening the reconstruction bottleneck that is essential for anomaly-normal separability. Second, DiT surpasses CLIP [19] under the frozen setting, benefiting from its deeper architecture and higher-dimensional representation, which provide richer semantic capacity. Additionally, the pre-training model as a robust encoder backbone endows it with inherent robustness to distribution shift.

Table 5: Ablation on semantic encoder. The best performance is in bold.

S_{enc}	D_e / d_s	Frozen	Image-level			Pixel-level		
			AUROC	AP	$F1_{max}$	AUROC	AP	$F1_{max}$
CLIP	12 / 768		99.2	99.4	98.4	98.0	69.5	69.8
CLIP	12 / 768	✓	99.6	**99.9**	99.1	98.5	69.8	70.1
DiT	28 / 1152		99.5	99.8	99.0	98.5	70.9	70.0
DiT	28 / 1152	✓	**99.7**	**99.9**	**99.3**	**98.6**	**71.3**	**70.7**

Table 6: Backbone encoder attribution study. The best performance is in bold.

Enc	Framework	Image-level			Pixel-level		
		AUROC	AP	$F1_{max}$	AUROC	AP	$F1_{max}$
DINOv2-L	Native	99.5	99.3	99.1	98.4	69.5	69.1
DINOv3-L	Native	99.6	99.8	99.0	98.4	70.4	69.7
DINOv2-L	HLRAD	99.6	**99.9**	99.2	98.5	69.8	70.5
DINOv3-L	HLRAD	**99.7**	**99.9**	**99.3**	**98.6**	**71.3**	**70.7**

Base Encoder. As shown in Tab. 6, we compare HLRAD with a native reconstruction version baseline equipped with MLP projection, dropout and a linear-attention decoder. The results show that HLRAD consistently achieves better performance. Notably, some DINOv2-based methods [7] report that replacing DINOv2 [17] with DINOv3 [22] can even degrade some metrics. In contrast, HLRAD does not suffer from this issue, suggesting that our framework more effectively exploits deep semantic representations rather than relying solely on a stronger frozen encoder.

5 Conclusion

In this paper, we propose HLRAD, a unified anomaly detection framework that challenges the prevailing paradigm of latent space compression by introducing dimensionality expansion for feature reconstruction. HLRAD encodes features into a higher-dimensional latent space via a semantic branch and a dimension-expanding MLP. To further enhance normal pattern modeling, the Selective Gated Fusion mechanism adaptively embeds high-dimensional semantic features to salient spatial positions, ensuring the deep semantic reconstruction learning. Extensive experiments on three classic industrial anomaly detection datasets demonstrate that HLRAD achieves outstanding results in the unified setting.

Discussion and Limitations. HLRAD relies on extracting reliable semantic representations from normal samples. When the dataset contains more categories and more complex normal variations, modeling the semantic distribution of each category becomes more difficult under a unified task setting. As a result, some normal patterns may be insufficiently represented, which can weaken image-level anomaly discrimination. This is reflected on Real-IAD, where the larger category diversity leads to weaker image-level results. In future work, we will explore more adaptive semantic modeling to improve robustness in a larger scale unified anomaly detection situation.

References

1. Bao, J., Sun, H., Deng, H., He, Y., Zhang, Z., Li, X.: BMAD: Benchmarks for medical anomaly detection. In: CVPRW. pp. 4042–4053 (2024)
2. Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., Steger, C.: The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. IJCV **129**(4), 1038–1059 (2021)
3. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9737–9746 (June 2022)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020), https://arxiv.org/abs/2010.11929
5. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR. pp. 12873–12883 (2021)
6. Guo, J., Lu, S., Jia, L., Zhang, W., Li, H.: ReContrast: Domain-specific anomaly detection via contrastive reconstruction. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 10721–10740 (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/228b9279ecf9bbafe582406850c57115-Paper-Conference.pdf
7. Guo, J., Lu, S., Zhang, W., Chen, F., Li, H., Liao, H.: Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20405–20415 (June 2025)
8. He, H., Bai, Y., Zhang, J., He, Q., Chen, H., Gan, Z., Wang, C., Li, X., Tian, G., Xie, L.: MambaAD: Exploring state space models for multi-class unsupervised anomaly detection. In: NeurIPS. pp. 71162–71187 (2024)
9. He, H., Zhang, J., Chen, H., Chen, X., Li, Z., Chen, X., Wang, Y., Wang, C., Xie, L.: A diffusion-based framework for multi-class anomaly detection. In: AAAI. pp. 8472–8480 (2024)
10. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013), https://arxiv.org/abs/1312.6114
11. Liu, Z., Zhou, Y., Xu, Y., Wang, Z.: Simplenet: A simple network for image anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20402–20411 (June 2023)
12. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017). https://doi.org/10.48550/arXiv.1711.05101
13. Lu, R., Wu, Y., Tian, L., Wang, D., Chen, B., Liu, X., Hu, R.: Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. In: NeurIPS. pp. 18341–18353 (2023)
14. Luo, W., Cao, Y., Yao, H., Zhang, X., Lou, J., Cheng, Y., Shen, W., Yu, W.: Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9974–9983 (June 2025)
15. Mishra, P., Verk, R., Fornasier, D., Piciarelli, C., Foresti, G.L.: Vt-adl: A vision transformer network for image anomaly detection and localization. In: 30th IEEE/IES International Symposium on Industrial Electronics (ISIE) (2021)

16. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: NeurIPS (2017)
17. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023). https://doi.org/10.48550/arXiv.2304.07193
18. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022), arXiv:2212.09748v2
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
20. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
21. Roth, K., Pemberton, L., Debard, T., Akcay, S., et al.: Towards total recall in industrial anomaly detection. In: CVPR. pp. 14318–14328 (2022)
22. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: Dinov3. arXiv preprint arXiv:2508.10104 (2025). https://doi.org/10.48550/arXiv.2508.10104
23. Su, J., Murtadha, A., Lu, Y., Wen, B., Pan, S., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. arXiv preprint arXiv:2104.09864 (2021), arXiv:2104.09864v5
24. Wang, C., Zhu, W., Gao, B.B., Gan, Z., Zhang, J., Gu, Z., Qian, S., Chen, M., Ma, L.: Real-IAD: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In: CVPR. pp. 16418–16427 (2024)
25. Wei, X.S., Zhang, C.L., Li, Y., Xie, C.W., Wu, J., Shen, C., Zhou, Z.H.: Deep descriptor transforming for image co-localization. arXiv preprint arXiv:1705.02758 (2017). https://doi.org/10.48550/arXiv.1705.02758
26. Yao, X., Zhang, C., Li, R., Sun, J., Liu, Z.: One-for-all: Proposal masked cross-class anomaly detection. In: AAAI. pp. 4792–4800 (2023). https://doi.org/10.1609/aaai.v37i4.25604
27. You, Z., Cui, L., Shen, T., Cheng, K., Zhu, W., Li, N., Ma, L.: A unified model for multi-class anomaly detection. pp. 18971–18981 (2022)
28. Zhang, J., Chen, X., Wang, Y., Wang, C., Liu, Y., Li, X., Yang, M., Tao, D.: Exploring plain vit reconstruction for multi-class unsupervised anomaly detection. arXiv preprint **arXiv:2312.07495** (2023), https://arxiv.org/abs/2312.07495
29. Zhang, X., Li, S., Li, X., Huang, P., Shan, J., Chen, T.: Destseg: Segmentation guided denoising student-teacher for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3914–3923 (June 2023)
30. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025), https://arxiv.org/abs/2510.11690
31. Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: ECCV. pp. 392–408 (2022). https://doi.org/10.1007/978-3-031-19769-7_23