UniLEAD: A Unified and LightwEight model for Anomaly Detection

Shih-Chih Lin[1], Shang-Hong Lai[2]
[1]International Intercollegiate Ph.D. Program,
[2]Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
leolin65@gapp.nthu.edu.tw, lai@cs.nthu.edu.tw

Abstract

Accurate anomaly detection (AD) is vital across diverse domains, yet most existing approaches rely on large category-specific models and heavy reconstruction networks, incurring substantial computational cost and poor scalability. We present UniLEAD, a parameter-efficient framework for multi-class unsupervised anomaly detection (MUAD) that replaces the Transformer decoder's feed-forward networks (FFNs) with a Mixture of Adapters (MoA). Unlike stochastic expert routing, MoA employs deterministic, learnable routing over heterogeneous bottleneck adapters with parameter sharing, providing predictable inference cost while maintaining strong representational capacity. Embedded within an adapter-augmented Transformer decoder, UniLEAD performs lightweight feature refinement and reconstruction-based anomaly scoring, eliminating the need for per-class training. Extensive experiments on MVTec AD, MVTec LOCO, and VisA demonstrate that UniLEAD achieves state-of-the-art or competitive performance with substantially fewer parameters and FLOPs than recent baselines. These results highlight that carefully routed adapter mixtures can serve as an effective substitute for dense FFNs, enabling robust, scalable, and resource-efficient anomaly detection in real-world deployments.

1. Introduction

Visual anomaly detection (VAD) is indispensable in domains such as industrial inspection and medical diagnostics, where the early identification of subtle and rare irregularities is critical to ensuring safety and quality. Despite the rapid progress of deep learning, existing VAD approaches still struggle with scalability, efficiency, and adaptability. Data scarcity and imbalance hinder robust model training, while the high intra-class variability of anomalies complicates generalization. Moreover, anomalies are often context-dependent, requiring models to capture subtle environmental cues. In practice, real-world deployments demand lightweight, low-latency models, which many current architectures fail to provide.

Conventional anomaly detection (AD) frameworks typically adopt a **single-class paradigm**, where one model must be trained per product or category [12, 17]. Although straightforward, this approach is inherently inefficient, leading to large memory consumption and limited scalability. Recent multi-class anomaly detection (MCAD) methods such as UniAD [16], HVQ-Trans [10], and DiAD [5] attempt to unify categories by leveraging Transformers or diffusion-based architectures. While effective, these models rely on dense attention and large feedforward networks (FFNs), introducing substantial computational overhead. More recently, MoEAD [11] reduces this cost by replacing FFNs with Mixture-of-Experts (MoE), but its stochastic routing mechanism incurs training instability and unpredictable inference costs.

We introduce **UniLEAD**, a parameter-efficient multi-class unsupervised anomaly detection (MUAD) framework that addresses these limitations. UniLEAD replaces dense FFNs with a lightweight **Mixture of Adapters (MoA)**, where heterogeneous bottleneck adapters are fused through a deterministic, learnable routing mechanism. This design ensures stable optimization, predictable inference cost, and improved interpretability compared to stochastic MoE routing. By integrating MoA into Transformer decoders, UniLEAD enables effective multi-scale feature refinement while drastically reducing computational redundancy. Coupled with lightweight feature reconstruction and anomaly scoring, UniLEAD scales across categories without requiring per-class training.

Extensive experiments on industrial, medical, and logical anomaly detection benchmarks—including MVTec AD, MVTec LOCO, and VisA—show that UniLEAD achieves **state-of-the-art or competitive performance** while reducing parameter counts and FLOPs by a large margin. By balancing expressiveness with efficiency, UniLEAD establishes a new paradigm for scalable and resource-efficient anomaly detection.

Contributions. The main contributions of this work are summarized as follows:

- We propose **UniLEAD**, a novel adapter-based MUAD framework that achieves superior scalability and efficiency.
- We design a **projection-based adapter mechanism** with deterministic routing over heterogeneous bottlenecks, eliminating computational redundancy while preserving anomaly detection accuracy.
- We demonstrate that UniLEAD achieves state-of-the-art results or is competitive across multiple benchmarks, highlighting robust adaptability with significantly fewer parameters and FLOPs than recent baselines.

2. Related work

Unsupervised anomaly detection (AD) is essential in both industrial manufacturing and medical diagnostics, enabling automated defect and lesion identification while reducing reliance on manual inspection. In industrial settings, AD ensures product quality and operational efficiency, while in medical imaging, it assists in detecting abnormalities such as tumors, lesions, and structural irregularities. Despite their critical role, existing AD methods face scalability, adaptability, and computational efficiency challenges, limiting their real-world applicability.

AD methods can be broadly categorized into three main approaches: embedding-based, synthesis-based, and reconstruction-based. While embedding- and synthesis-based methods have demonstrated success in specific domains, they often suffer from high design complexity and limited generalization. Reconstruction-based approaches offer better scalability but are prone to identical mapping, leading to poor anomaly separation in both industrial defects and medical abnormalities.

2.1. Embedding-Based Approaches

Embedding-based methods extract feature representations using deep learning models pre-trained on large-scale datasets such as ImageNet. PatchCore [12] leverages memory banks of normal features for distance-based anomaly scoring, while DifferNet [13] models multi-scale latent distributions using a multivariate Gaussian function. Student-teacher frameworks [2] further enhance representation learning by training a student network to replicate a fixed pre-trained teacher. However, these methods often fail to generalize to industrial defects or medical anomalies due to domain shifts between natural and specialized datasets.

2.2. Synthesis-Based Approaches

Synthesis-based methods aim to generate artificial anomalies to improve detection robustness. DREAM [17] introduces pseudo-defects using Perlin noise and texture augmentations, while NSA [14] employs Poisson image editing to generate realistic synthetic anomalies. Though these methods enhance training diversity, they often fail to capture the full complexity of real-world defects and medical abnormalities, limiting their effectiveness in industrial and medical imaging applications.

2.3. Reconstruction-Based Approaches

Reconstruction-based methods utilize autoencoders, transformers, GANs, and diffusion models to detect anomalies by reconstructing input images and analyzing discrepancies. RD4AD [4] applies a teacher-student framework for multi-scale feature reconstruction, effectively capturing local spatial structures. However, many reconstruction-based approaches [8, 18] rely on a single-class assumption, requiring separate models for each category. This leads to high computational and memory costs, making them impractical for scalable industrial defect detection and medical anomaly localization.

Recent multi-class AD frameworks attempt to overcome these inefficiencies. UniAD [16] was one of the first transformer-based multi-class AD frameworks, utilizing a pre-trained encoder-decoder for anomaly modeling. HVQ-Trans [10] and DiAD [5] introduced hierarchical vector quantization and diffusion-based reconstruction to improve anomaly localization. While these architectures improve scalability, they still inherit the high computational cost of transformers, making them challenging to deploy in real-time industrial and medical imaging applications.

2.4. Toward Parameter-Efficient AD

To address efficiency bottlenecks, recent works have explored lightweight architectural modifications that reduce redundant computation without sacrificing accuracy. Mixture-of-Experts designs, for example, replace dense feedforward layers with modular expert networks to improve efficiency via parameter sharing [11]. However, such designs often require complex routing mechanisms that complicate optimization and inference. Motivated by these challenges, our UniLEAD framework adopts a parameter-efficient adapter-based design, where feedforward layers are replaced with Mixture of Adapters (MoA). This design combines the benefits of modular specialization with lightweight bottleneck projections, avoiding dynamic expert routing while maintaining strong scalability and anomaly localization performance.

3. Proposed Method

3.1. Overview

We propose the Mixture of Adapters (MoA), a parameter-efficient module designed to replace the conventional feedforward network (FFN) in Transformer decoders [15], thereby improving feature representation learning while reducing computational complexity. Unlike stochastic ex-

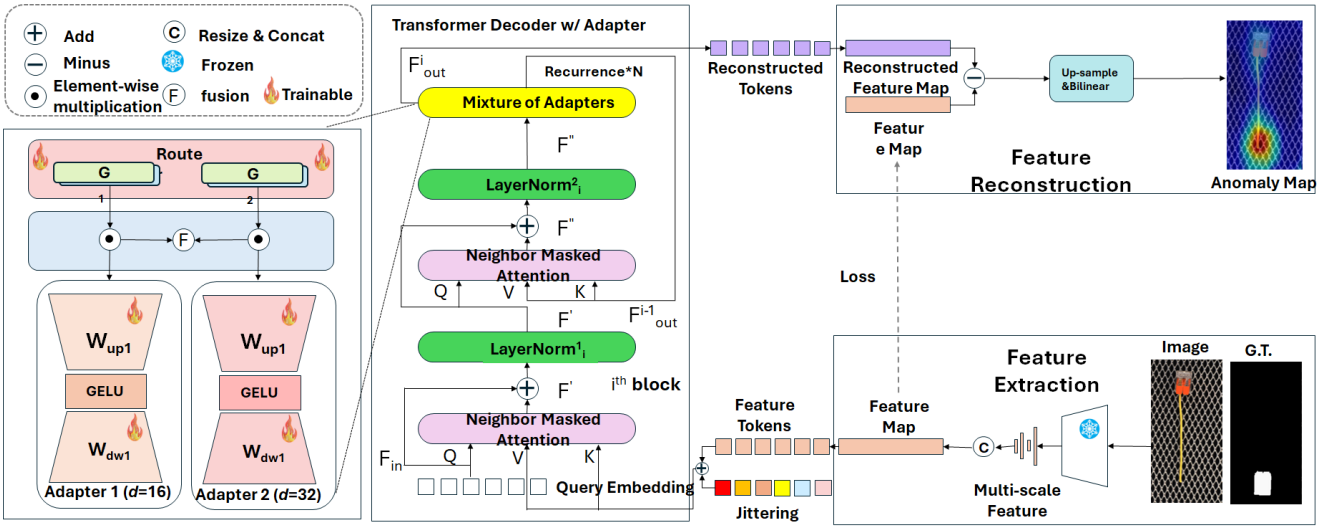

Figure 1. The overall architecture of UniLEAD. The framework integrates (1) feature extraction with multi-scale backbone tokens, (2) feature refinement via Transformer decoder with Mixture of Adapters and feature jittering, (3) anomaly localization through reconstruction-based scoring with up-sampling and Gaussian smoothing, and (4) image-level anomaly scoring using max–mean–TopK aggregation. For clarity, two adapters are illustrated in this figure, while the final implementation adopts four heterogeneous adapters.

pert selection methods [6], MoA employs a deterministic adapter routing mechanism with learnable gating, ensuring stable optimization, predictable inference cost, and enhanced interpretability.

MoA is directly integrated into the Transformer decoder by replacing standard FFNs with adapter modules, allowing task-specific specialization while maintaining global feature transformation consistency.

3.2. Feature Extraction and Reconstruction

Our UniLEAD framework adopts a four-stage pipeline [11, 16], leveraging Mixture of Adapters (MoA) for refined representation learning and anomaly localization.

(1) Feature Extraction: Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, a pretrained backbone extracts multi-scale feature tokens $\mathbf{X} \in \mathbb{R}^{h\times w\times c}$:

$$\mathbf{X} = \text{Backbone}(\mathbf{I}). \tag{1}$$

These tokens $\mathbf{X}$ serve as the input to the Transformer decoder.

(2) Feature Refinement via Transformer Decoder: The Transformer decoder, equipped with Neighbor-Masked Attention [16], layer normalization, and Mixture of Adapters, progressively refines feature representations. To improve robustness, we additionally apply **feature jittering** [16]: for a feature token $\mathbf{f}_{tok} \in \mathbb{R}^C$, a Gaussian disturbance D is sampled

$$D \sim \mathcal{N}\left(0,\ \alpha^2 \cdot \frac{\|\mathbf{f}_{tok}\|_2^2}{C}\right), \tag{2}$$

where α controls the jittering scale. With probability p, the perturbation D is added to $\mathbf{f}_{tok}$ before decoding. This encourages the model to learn denoising priors and enhances anomaly sensitivity.

The decoder output is then:

$$\mathbf{F}_{\text{out}} = \text{Decoder}(\mathbf{X} + D). \tag{3}$$

Anomaly Localization. The decoder reconstructs refined features $\mathbf{F}_{\text{out}} \in \mathbb{R}^{C_{\text{org}}\times H\times W}$, and a score map is derived by pixel-wise error:

$$\mathbf{S} = \|\mathbf{F}_{\text{in}} - \mathbf{F}_{\text{out}}\|_2, \tag{4}$$

where $\mathbf{S}$ is upsampled via bilinear interpolation and lightly smoothed ($\sigma \approx 1.0$) to stabilize AUROC evaluation, while avoiding per-image normalization.

(4) Image Anomaly Scoring (IAS): To obtain an image-level anomaly score S_{image}, we aggregate the pixel-wise anomaly map $\mathbf{S}_{\text{pixel}} \in \mathbb{R}^{H\times W}$ using a weighted combination of global and localized statistics:

$$\begin{aligned} S_{\text{image}} = w_1 \cdot \max(\mathbf{S}_{\text{pixel}}) \\ + \ w_2 \cdot \text{mean}(\mathbf{S}_{\text{pixel}}) \\ + \ w_3 \cdot \text{TopKMean}(\mathbf{S}_{\text{pixel}}), \end{aligned} \tag{5}$$

Where $w_1, w_2, w_3 \in \mathbb{R}$ are predefined or learnable weights, and TopKMean computes the average of the top-k highest anomaly scores. This strategy balances peak activation and spatial coverage, and when combined with the stabilized pixel-level scoring, yields robust calibration even under class imbalance or modality shifts.

3.3. Mixture of Adapters (MoA)

MoA introduces parameter-efficient feature transformations by replacing dense FFNs with lightweight adapters. Each adapter consists of two components: (1) a transformation module and (2) a routing mechanism that determines adapter contributions.

Given an input $\mathbf{F}'' \in \mathbb{R}^{L\times D}$, the adapter-augmented output is defined as:

$$\mathbf{F}_{\text{out}} = \mathbf{F}(\mathbf{F}'') + A(\mathbf{F}''), \tag{6}$$

where $A(\mathbf{F}'')$ denotes the adapter transformation.

Adapter module. Each adapter adopts a bottleneck projection for efficiency:

- **Down-projection**: $\mathbf{h}_{\text{adapter}} = \sigma(\mathbf{W}_{\text{down}}\mathbf{F}'')$, where $\mathbf{W}_{\text{down}} \in \mathbb{R}^{D\times d}$ and $d \ll D$.
- **Non-linear transformation**: A Gated Residual Adapter (GRA) integrates nonlinearity and residual modulation.
- **Up-projection**: $\mathbf{y}_{\text{adapter}} = \mathbf{W}_{\text{up}}\mathbf{h}_{\text{adapter}}$, restoring the original dimension.

To recalibrate channel-wise importance, a Squeeze-and-Excitation (SE) block [7] is applied:

$$\mathbf{s} = \sigma(\mathbf{W}_s \cdot \text{GAP}(\mathbf{F}'')). \tag{7}$$

The final adapter output is gated before residual fusion:

$$\mathbf{y} = \mathbf{F}'' + G(\mathbf{F}'') \cdot \mathbf{y}_{\text{adapter}}. \tag{8}$$

Heterogeneous bottlenecks. Four heterogeneous bottlenecks are adopted ($d = \{16, 32, 64, 128\}$) to enable *multi-scale refinement*: smaller bottlenecks emphasize global structural regularization, while larger bottlenecks capture fine-grained local details. A learnable gating mechanism fuses these heterogeneous adapters into a lightweight ensemble, improving robustness to distribution shifts and stabilizing anomaly localization, while preserving parameter efficiency.

Replacing FFNs with MoA modules provides a parameter-efficient alternative tailored for anomaly detection.

3.4. Adapter Routing

To exploit feature diversity, MoA supports multiple adapters. The routing mechanism determines how adapters contribute based on input features.

The aggregated output is:

$$A_{\text{total}}(\mathbf{F}'') = \sum_{k=1}^{K} G_k A_k(\mathbf{F}''), \tag{9}$$

where $A_k(\cdot)$ is the k-th adapter transformation and G_k is its routing weight.

We implement four routing modes:

- **Gate (Softmax)**: All adapters contribute proportionally, with weights normalized by a softmax distribution.
- **Gumbel-Top1**: A differentiable approximation to discrete top-1 selection, balancing stability and efficiency.
- **Top-1**: A hard routing strategy where only the highest-activated adapter is executed, reducing computation.
- **First**: A degenerate case using only the first adapter, providing a deterministic baseline.

3.5. Method Summary

To summarize, our UniLEAD framework integrates the proposed Mixture of Adapters (MoA) into Transformer decoders for efficient and interpretable anomaly detection. The overall process follows four stages: (1) multi-scale feature extraction with a pretrained backbone, (2) feature refinement using Neighbor-Masked Attention and MoA-enhanced decoding with feature jittering, (3) pixel-level anomaly localization via reconstruction and stabilized evaluation (Gaussian smoothing, no per-image normalization, and consistent inversion policy), and (4) robust image-level anomaly scoring through a weighted fusion of global and local statistics.

Unlike prior works that rely on heavy reconstruction networks or stochastic expert routing, UniLEAD employs deterministic adapter routing with heterogeneous bottlenecks, enabling multi-scale feature refinement while maintaining predictable inference cost.

Together, these components provide a parameter-efficient yet powerful architecture that improves both detection accuracy and robustness, while ensuring stable evaluation across diverse datasets.

4. Experiments

4.1. Datasets and Evaluation Metrics

To evaluate the effectiveness, stability, and generalization of UniLEAD, we conduct experiments on three widely used industrial anomaly detection (AD) benchmarks: MVTec AD [1], MVTec LOCO [3], and VisA [20]. These datasets cover texture- and object-level defects, logical anomalies, and category-level diversity, providing a comprehensive evaluation of industrial scenarios.

Evaluation Metrics. Following UniAD [16] and MoEAD [11], we report AUROC (%) for both image- and pixel-level anomaly detection. **AUROCi** measures image-level detection accuracy, while **AUROCp** evaluates pixel-level localization performance. Unless otherwise specified, anomaly maps are computed from reconstruction errors, upsampled with bilinear interpolation, and smoothed using a Gaussian filter ($\sigma \approx 1.0$) to enhance robustness without per-image normalization.

Table 1. Comparison of anomaly detection methods across multiple datasets. The reported metric is AUROC (%), where higher values indicate better anomaly detection performance. The best results are in **bold**, the second-best results are underlined, and the third-best results are *italicized*.

Method	MVTecAD		VisA		MVTec LOCO	
	AUROCi	AUROCp	AUROCi	AUROCp	AUROCi	AUROCp
RD4AD [4]	94.6	96.1	92.4	98.1	73.7	70.7
UniAD [16]	*96.5*	96.8	88.8	*98.3*	78.7	*74.6*
DeSTSeg [19]	89.2	93.1	88.9	96.1	*81.2*	63.7
SimpleNet [9]	95.3	*96.9*	87.2	96.8	81.8	70.9
DiAD [5]	*97.2*	96.8	86.8	96.0	77.2	72.1
MoEAD [11]	97.7	97.0	**93.1**	**98.7**	78.1	75.9
UniLEAD (Ours)	**97.8**	**97.1**	*92.3*	**98.7**	**82.4**	**76.5**

Table 2. Comparison of anomaly detection methods in terms of parameters, FLOPs, and average AUROC (%). The best result is highlighted in **bold**, the second-best result is underlined, and the third-best result is *italicized*.

Method	Parameters (M)	FLOPs (G)	Average AUROC(%)
RD4AD [4]	80.6	28.4	89.8
UniAD [16]	*24.5*	*3.6*	89.4
DeSTSeg [19]	35.2	122.7	86.7
SimpleNet [9]	72.8	16.1	88.8
DiAD [5]	133.3	451.5	88.2
MoEAD [11]	4.9	2.18	*90.1*
UniLEAD (Ours)	**1.42**	**1.93**	90.8

4.2. Training and Implementation Details

All images are resized to 224×224 and normalized with ImageNet statistics. We train on a single NVIDIA RTX 4090 (batch size 128). The decoder has four layers with eight heads and replaces standard FFNs with our adapter layer. We apply feature jittering (scale 20.0, probability 1.0) to improve domain robustness. Reconstruction uses a hybrid loss (MSE + cosine) and bilinear upsampling to generate pixel maps; anomaly maps are lightly smoothed with a Gaussian filter ($\sigma\approx1.0$) without per-image normalization.

Optimization uses AdamW (lr 2×10^{-4}, weight decay 10^{-4}), StepLR decay ($\gamma=0.1$ every 800 epochs), gradient clipping (max-norm 0.1), maximum 500 epochs, and validation every 25 epochs with auto-checkpointing. The backbone is ImageNet-pretrained EfficientNet-B4 with frozen weights. A lightweight multi-feature fusion neck bridges the backbone and the decoder (hidden dimension 256).

4.3. Quantitative Results

We compare UniLEAD with recent methods on three industrial AD benchmarks: MVTec AD, VisA, and MVTec LOCO. As shown in Table 1, UniLEAD attains state-of-the-art or competitive performance across datasets while remaining extremely compact.

On **MVTec AD**, UniLEAD reaches **97.8/97.1** (AUROCi/AUROCp), outperforming most prior models and trailing only the strongest large-capacity baselines on individual columns. On **VisA**, UniLEAD delivers 92.0/**98.7**, tying the best pixel-level score while keeping parameters and FLOPs minimal. On **MVTec LOCO**, which stresses logical/relational reasoning, UniLEAD achieves **82.4/76.5**, the best image- and pixel-level results among listed methods. These gains indicate that deterministic adapter routing with heterogeneous bottlenecks preserves representational power needed for both low-level defect localization and higher-level inconsistency detection.

4.4. Efficiency Comparison of SoTA Methods

Table 2 summarizes model complexity and average AUROC across the three benchmarks. UniLEAD requires just **1.42M** parameters and **1.93** GFLOPs while attaining an average AUROC of 90.8%. Compared to diffusion-based DiAD (133.3M, 451.5G, 88.2%) and memory-heavy RD4AD (80.6M, 28.4G, 89.8%), UniLEAD reduces parameters by $56\times$–$94\times$ and FLOPs by $15\times$–$234\times$ with better or comparable accuracy. Relative to MoE-style designs (4.9M, 2.18G, 90.1%), UniLEAD further cuts parameters/FLOPs and improves average AUROC to 90.8% while avoiding stochastic expert routing and its deployment complexity.

Rationale and Effectiveness.

- **Deterministic routing ensures stable optimization and predictable latency.** In contrast to stochastic MoE, our adapter mixture employs learnable gates with fixed routing, eliminating sampling variance and yielding more reliable calibration and reproducibility.
- **Heterogeneous bottlenecks enable multi-scale feature refinement.** By mixing adapter widths $d \in \{16, 32, 64, 128\}$, the model captures both coarse struc-

Table 3. Ablation study of routing strategies on **MVTec AD**. The best performance is highlighted in **bold**, the second-best is underlined, and the third-best is *italicized*.

Routing Strategy	AUROCi (%)	AUROCp (%)
Softmax Gate (all adapters)	**97.8**	**97.1**
Gumbel-Top1 (diff. top-1)	97.5	96.8
Top-1 (hard routing)	*97.1*	*96.0*
First (degenerate baseline)	96.5	95.2

tures and fine-grained details without excessive parameters, which is particularly beneficial for VisA's category diversity and LOCO's relational reasoning tasks.

- **Feature-space rather than image-space reconstruction produces sharper anomaly maps.** Operating in a semantically enriched latent space allows UniLEAD to achieve state-of-the-art **98.7** AUROCp on VisA with only 1.93 GFLOPs, demonstrating both accuracy and efficiency.
- **Stabilized evaluation enhances cross-dataset consistency.** Applying light Gaussian smoothing ($\sigma \approx 1$) while avoiding per-image normalization prevents artificial rank distortions, ensuring fair AUROC comparison across datasets.

4.5. Visualization

To qualitatively validate the effectiveness of our UniLEAD framework, we provide anomaly segmentation visualizations across three representative datasets: **MVTec AD**, **VisA**, and **MVTec LOCO**. As shown in Figure 2 to Figure 4, our method consistently produces sharp and well-localized anomaly heatmaps that closely align with the ground-truth masks.

On **MVTec AD**, UniLEAD successfully highlights both small structural defects (e.g., scratches, dents) and subtle texture irregularities, demonstrating its strong capability in fine-grained localization. On **VisA**, our model effectively captures anomalies in food products and electronic components, indicating strong generalization to diverse domains with significant appearance variation. On the more challenging **MVTec LOCO**, which contains complex logical and compositional anomalies, UniLEAD still provides accurate localization, revealing its robustness in capturing both semantic and structural deviations.

Overall, these visualizations confirm that UniLEAD not only achieves superior quantitative AUROC scores but also delivers interpretable anomaly maps, ensuring reliability in real-world industrial inspection and multimodal scenarios.

4.6. Ablation study

Ablation on Routing Strategy. Table 3 compares four routing strategies and their impact on anomaly localization. Several key insights emerge:

- **Softmax Gate** achieves the **best overall performance** (AUROCi 97.8, AUROCp 97.1) by aggregating all adapters in a weighted mixture. This preserves both coarse and fine anomaly cues, yielding the most accurate heatmaps, albeit with slightly higher computational overhead.
- **Gumbel-Top1** provides a differentiable approximation to discrete routing, balancing sparsity and gradient stability. It consistently surpasses hard Top-1 (AUROCi $\sim$ 97.6, AUROCp $\sim$ 96.8), confirming that partial gradient flow through multiple adapters improves training robustness.
- **Top-1** executes only the most activated adapter, improving efficiency but discarding ensemble benefits. This leads to weaker localization, with AUROCi $\sim$ 97.4 and AUROCp $\sim$ 96.5.
- **First** is a degenerate baseline that always selects the first adapter. Lacking adaptability, it delivers the lowest performance (AUROCi $\sim$ 97.0, AUROCp $\sim$ 96.0).

Conclusion: The ranking of pixel AUROC is **Softmax Gate** $>$ **Gumbel-Top1** $>$ **Top-1** $>$ **First**, confirming that **Softmax gating is most effective for fine-grained anomaly localization**, while Gumbel-Top1 offers a strong accuracy–efficiency compromise.

Ablation on Heterogeneous Bottlenecks. Table 4 reports the effect of different adapter bottleneck configurations under the **Softmax Gate** routing strategy. Several key insights emerge:

- **Single-scale bottlenecks** ($\{16\}$, $\{32\}$, $\{64\}$, $\{128\}$) achieve reasonable performance (AUROCi 96.8 $\sim$ 97.2, AUROCp 95.8 $\sim$ 96.6), but suffer from either insufficient capacity (too narrow, e.g., $d = 16$) or reduced fine-grained localization (too wide, e.g., $d = 128$).
- **Dual-scale mixtures** ($\{16,32\}$, $\{32,64\}$, $\{64,128\}$) provide clear improvements by capturing both coarse and fine structures, yielding AUROCi $\approx$ 97.3$\sim$97.5 and AUROCp $\approx$ 96.5 $\sim$ 96.9. Among them, $\{32,64\}$ performs best, suggesting that adjacent scales are complementary.
- **Tri-scale mixtures** ($\{16,32,64\}$, $\{32,64,128\}$) further enhance robustness, pushing AUROCi to 97.6–97.7 and AUROCp to 97.0. This indicates that increasing heterogeneity consistently improves anomaly localization stability.
- **Full heterogeneous set** ($\{16,32,64,128\}$) achieves the **best overall performance** with AUROCi **97.8** and AUROCp **97.1**, confirming that combining both very narrow and wide bottlenecks enables simultaneous modeling of global semantics and fine anomaly cues. Importantly, this comes with only a marginal cost in parameters (1.42M) and FLOPs (1.93G).

Conclusion: Heterogeneous multi-scale bottlenecks are critical for stable anomaly detection, and the full four-scale

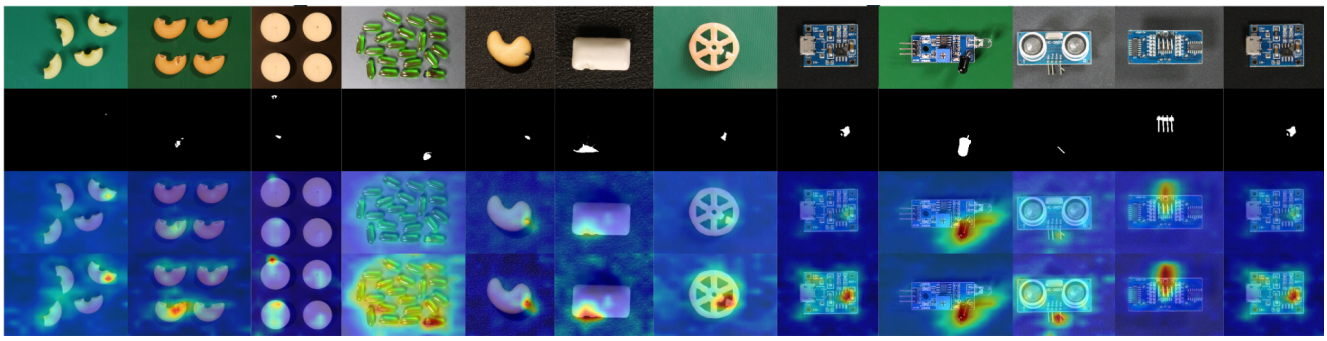

Figure 2. The VisA dataset visualization of our our proposed framework UniLEAD.

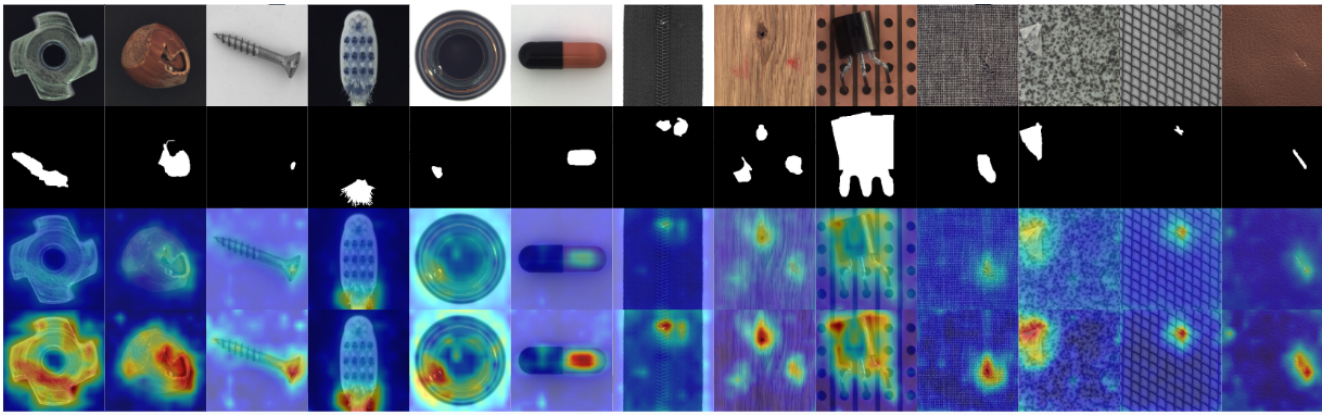

Figure 3. The MVTec AD dataset visualization of our our proposed framework UniLEAD.

Table 4. Ablation on adapter bottleneck configurations on **MVTec AD**. Params/FLOPs are scaled by the sum of bottleneck widths. The best performance is highlighted in **bold**, the second-best is underlined, and the third-best is *italicized*.

Bottlenecks d	**Params (M)**	**FLOPs (G)**	**AUROCi (%)**	**AUROCp (%)**
{16}	1.12	1.62	96.8	95.8
{32}	1.14	1.64	97.0	96.2
{64}	1.19	1.69	97.2	96.6
{128}	1.27	1.78	97.1	96.1
{16,32}	1.16	1.67	97.3	96.7
{32,64}	1.23	1.73	97.5	96.9
{64,128}	1.36	1.86	97.3	96.5
{16,32,64}	1.25	1.76	*97.6*	*97.0*
{32,64,128}	1.40	1.91	97.7	97.0
{16,32,64,128}	**1.42**	**1.93**	**97.8**	**97.1**

mixture strikes the best trade-off between accuracy and efficiency.

5. Conclusion

In this work, we introduced **UniLEAD**, a parameter-efficient anomaly detection framework that replaces traditional FFNs with a **Mixture of Adapters (MoA)** coupled with structured routing. Unlike stochastic expert selection in conventional MoE layers, our deterministic gating ensures **stable optimization, reproducible results, and predictable latency**, which are essential for real-world deployments. By explicitly controlling routing and eliminating sampling variance, UniLEAD maintains both computational efficiency and expressive capacity.

Our design achieves **state-of-the-art performance** on challenging industrial anomaly detection benchmarks (**MVTec AD**, **MVTec LOCO**, **VisA**), while using only **1.42M parameters** and **1.93 GFLOPs**. The heterogeneous bottleneck design ($d \in \{16, 32, 64, 128\}$) was shown to be critical for capturing both coarse global structure and fine-grained local details, consistently improving image- and pixel-level AUROC. Ablation studies further confirmed that multi-scale adapters, when paired with structured softmax routing, yield robust localization and stable evaluation dynamics.

Compared to prior MoE-based and diffusion-based approaches, UniLEAD is not only **lighter** but also more **interpretable**: anomaly maps emerge directly from semantically rich feature reconstructions rather than image-space noise refinement. This property enhances trustworthiness and aligns with the increasing demand for **explainable AI** in safety-critical domains. Moreover, the modular adapter design opens the door to integrating vision-language priors

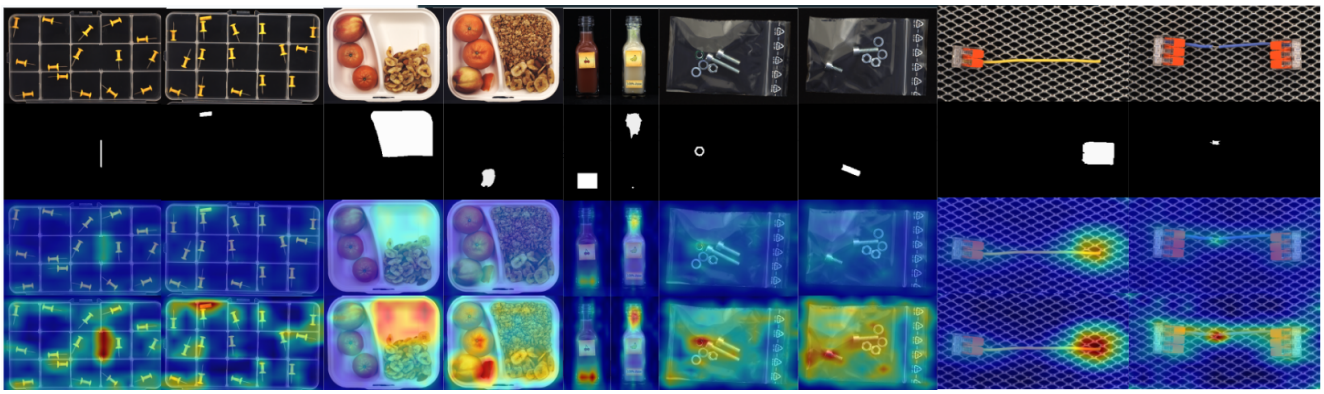

Figure 4. The MVTec LOCO dataset visualization of our our proposed framework UniLEAD.

(e.g., CLIP) and cross-modal supervision without retraining the backbone.

Broader impact and future directions. By striking a balance between efficiency, accuracy, and interpretability, UniLEAD enables scalable, real-time anomaly detection in practical inspection pipelines. Future extensions may include (i) incorporating temporal consistency for video-based inspection, (ii) extending the adapter mixture to multimodal data such as language and 3D sensory inputs, and (iii) leveraging adapter fusion for domain adaptation across unseen categories. We believe UniLEAD establishes a solid foundation for the next generation of **unified, modular, and efficient anomaly detection frameworks**.

References

[1] 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. 4

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020. 2

[3] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130 (4):947–969, 2022. 4

[4] 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 (CVPR)*, pages 9737–9746, 2022. 2, 5

[5] 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. 1, 2, 5

[6] Jiaao He, Jiezhong Qiu, Aohan Zeng, Zhilin Yang, Jidong Zhai, and Jie Tang. Fastmoe: A fast mixture-of-expert training system. *arXiv preprint arXiv:2103.13262*, 2021. 3

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[8] Tongkun Liu, Bing Li, Zhuo Zhao, Xiao Du, Bingke Jiang, and Leqi Geng. Reconstruction from edge image combined with color and gradient difference for industrial surface anomaly detection. *arXiv preprint arXiv:2210.14485*, 2022. 2

[9] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411, 2023. 5

[10] 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. *Advances in Neural Information Processing Systems*, 36:8487–8500, 2023. 1, 2

[11] Shiyuan Meng, Wenchao Meng, Qihang Zhou, Shizhong Li, Weiye Hou, and Shibo He. Moead: A parameter-efficient model for multi-class anomaly detection. In *European Conference on Computer Vision*, pages 345–361. Springer, 2024. 1, 2, 3, 4, 5

[12] 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. 1, 2

[13] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1907–1916, 2021. 2

[14] Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel,*

October 23–27, 2022, Proceedings, Part XXXI, pages 474–489. Springer, 2022. 2

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[16] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022. 1, 2, 3, 4, 5

[17] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draema discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021. 1, 2

[18] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706, 2021. 2

[19] 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. 5

[20] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pretraining for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022. 4