One Language-Free Foundation Model Is Enough for Universal Vision Anomaly Detection

Bin-Bin Gao
Tencent YouTu Lab, Shenzhen, China
csgaobb@gmail.com
Code and Models: https://github.com/gaobb/UniADet

Abstract

Universal vision anomaly detection (AD) aims to identify anomalous images and segment anomalous regions in open and dynamic scenarios, typically adhering to zero- and few-shot paradigms without dataset-specific fine-tuning. Recently, the field has seen significant progress driven by the integration of vision-language foundation models. However, we observe that current methods often struggle with laborious prompt engineering, elaborate adaptation modules, and complex training strategies that ultimately constrain their flexibility and generalizability. In this paper, we rethink the fundamental mechanism of vision-language models for AD and present UniADet, an embarrassingly simple, effective and general framework for Universal vision Anomaly Detection. Our approach is built on two key insights: First, we reveal that the primary function of the language encoder is merely to derive decision weights and we demonstrate that it is unnecessary, as these weights can be learned more directly and efficiently. Second, to resolve learning conflicts arising from disparate feature manifolds, we introduce a systematic decoupling strategy that learns independent weights across both distinct tasks (classification vs. segmentation) and hierarchical features. UniADet is highly simple and efficient, learning only decoupled weights and requiring 0.02M learnable parameters. It is inherently versatile, adapting seamlessly to various foundation models (*e.g.*, CLIP, DINOv2, and DINOv3). Extensive evaluations on 14 real-world benchmarks in the industrial and medical domains demonstrate that UniADet not only outperforms state-of-the-art zero- and few-shot methods by a substantial margin, but also outperforms full-shot AD methods for the first time. This empirical evidence highlights the potential of language-free frameworks for universal vision anomaly detection.

1 Introduction

Visual Anomaly Detection (AD) is a fundamental task in computer vision with critical applications ranging from industrial quality inspection [2, 58, 47] and medical imaging diagnostics [25, 19, 46, 6, 45] to hybrid subspecies discovery in biological sciences [39]. Traditional AD paradigms [38, 50, 12, 18, 14] typically necessitate extensive domain-specific training on normal images to model the distribution of normality. However, this requirement significantly hampers scalability and adaptability in open-world environments and privacy-constrained scenarios. Consequently, recent research has shifted toward zero-shot and few-shot AD, leveraging the robust zero-shot generalization capabilities of large-scale vision-language models (VLMs, *e.g.*, CLIP [37]).

Over the past two years, many CLIP-based zero-shot [22, 55, 4, 36, 33, 35, 15] and few-shot [57, 29] ADs have emerged, achieving impressive performance across diverse benchmarks. Despite their success, these methods often introduce significant architectural overhead, including intricate prompt

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

engineering, elaborate adapter modules, or convoluted training strategies (see Fig. 1(a)). These dependencies not only increase computational complexity but also constrain AD method to a textual modality, which potentially limits its compatibility with emerging vision-only foundation models such as DINOv3 [43]. This observation motivates a fundamental, yet overlooked, question: Is a vision-language model strictly necessary for universal (both zero- and few-shot) anomaly detection?

MetaUAS [13] recently demonstrated that vision-only models can achieve competitive one-shot performance through training on synthetic images. Extending this capability to a zero-shot setting remains challenging. We argue that visual perception should ideally function independently of language. By rethinking the mechanics of vision-language-based AD, we observe that both manual and learnable prompt embeddings essentially serve as weights or prototypes for anomaly classification and segmentation. We hypothesize that directly optimizing these weights in the visual latent space is equivalent to, and potentially more efficient than, deriving them via a text encoder. This shift enables the deployment of universal AD using vision-only foundation models like DINOv2 [7] and DINOv3 [43], significantly reducing the inductive bias imposed by textual descriptions.

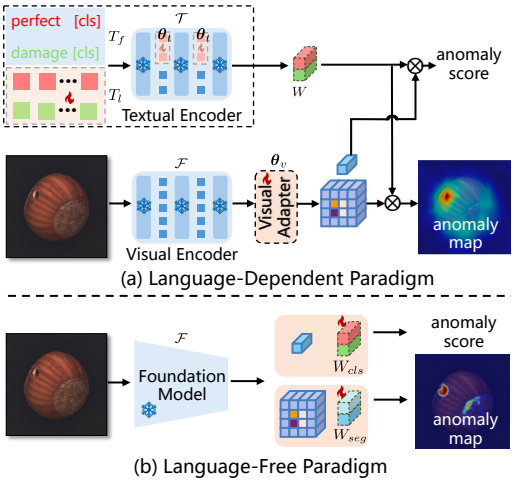


Figure 1: Comparison of current language-dependent and our language-free paradigm for zero-shot vision anomaly detection. (a) The language-dependent paradigm requires a heavy dual-branch pipeline, relying on a textual encoder and complex adapters ($\boldsymbol{\theta}_t$ and/or $\boldsymbol{\theta}_v$) to align visual features with fixed or learnable textual prompts. (b) The proposed language-free paradigm is *simple* (learning only task-related weights, W_{cls} and W_{seg}), *parameter-efficient* (about 0.015M learnable parameters), *general* (adapting a variety of foundation models, *e.g.*, CLIP and DINOv3), and *effective* (surpassing SOTA zero/few-shot and even full-shot AD methods).

Furthermore, we identify a critical limitation in existing VLM-based methods: the use of a shared weight matrix across multi-level hierarchical features. It is well-established that representations from different layers of a deep network reside on distinct manifolds. Moreover, global image tokens and local patch tokens exhibit disparate statistical distributions (see Fig. 3). Using a shared weight to adapt these heterogeneous features introduces a learning conflict, which compromises the model's discriminative power and generalization. To address this, we propose a Decoupled Anomaly Detector, which decouples global anomaly classification and local anomaly segmentation as distinct tasks with independent weights. This decoupling is extended across multi-scale feature hierarchies to preserve the unique geometric properties of each representation level. Our main contributions are summarized as follows:

- We rethink vision-language ADs and find that language prompts and encoders are unnecessary for zero-shot AD. This insight leads to an embarrassingly simple, parameter-efficient, and highly general language-free framework for universal anomaly detection.
- We fully decouple global anomaly classification and local anomaly segmentation across multi-scale hierarchical features, effectively mitigating the learning conflict between different feature manifolds and substantially improving AD performance.
- Comprehensive experiments on 6 industrial and 8 medical benchmarks demonstrate that our approach achieves state-of-the-art zero-shot and few-shot performance in both image-level classification and pixel-level segmentation. Notably, our 4-shot UniADet is the first to outperform the state-of-the-art full-shot methods on the large-scale Real-IAD benchmark.

2 Related Work

Unsupervised Anomaly Detection. Conventional unsupervised AD identifies anomalies by learning from a distribution of normal images. This field is primarily divided into three paradigms: *embedding-based* methods [8, 38, 40, 26] that leverage pre-trained features for density or distance estimation; *discrimination-based* methods [28, 51, 32] that reformulate AD as a supervised task via synthetic pseudo-anomalies; and *reconstruction-based* approaches [9, 50, 54, 48, 12, 18, 14] that assume that

anomalies yield higher residuals due to their absence from the training distribution. Despite their efficacy, these models are inherently constrained to seen classes and require expensive data collection and retraining for new scenarios, hindering rapid cross-domain adaptation.

Zero-Shot Anomaly Detection. Recent advances leverage Vision-Language Models (VLMs), particularly CLIP [37], to achieve zero-shot AD. WinCLIP [22] utilizes hand-crafted prompts with multi-scale windows for training-free detection. Subsequent works, such as AnomalyCLIP [55], AdaCLIP [4], VCP-CLIP [36], Bayes-PFL [35], and FAPrompt [56], focus on learning class-agnostic or abnormality-decomposed prompt embeddings. AA-CLIP [33], FE-CLIP [16] and AdaptCLIP [15] further introduce lightweight adapters for enhancing anomaly discrimination ability. However, we argue that these textual modifications introduce unnecessary complexity and may compromise CLIP's inherent representational robustness. While ACR [27], MuSc [30] and RareCLIP [20] explore batch-level information, they are less suited for privacy-sensitive environments. Our approach diverges by discarding the language branch entirely, pursuing a more streamlined and efficient architecture.

Few-Shot Anomaly Detection. Few-shot AD aims to adapt models using only a few normal images. While early methods [42, 21, 11, 10, 49] struggled to match unsupervised performance, recent approaches leveraging foundation models have set new benchmarks. WinCLIP+ [22] utilizes memory banks for retrieval, while InCtrl [57] integrates multi-level information for image-level classification. PromptAD [29] learns anomaly margins through normal textual prompts, whereas the recent FOUNDAD [52] trains a nonlinear projector via synthesized anomalies. Both PromptAD and FOUNDAD rely on few-shot normal images but require dataset-specific training. UniVAD [17] achieved state-of-the-art results by ensembling multiple models (*e.g.*, RAM [53], SAM [24], Grounding DINO [31], CLIP [37] and DINOv2 [7]). In contrast, we advocate for a training-free, single-model solution that maximizes computational efficiency without sacrificing performance.

3 Method

Problem Formulation: We aim to develop a universal anomaly detection framework capable of generalizing across diverse domains without target dataset-specific fine-tuning. This objective implies robust generalization to significant distribution shifts between base and target domains. Formally, let $\mathcal{D}_{\text{base}} = \{(X_i, Y_i, y_i)\}_{i=1}^{N}$ denote the base training set, where $X_i \in \mathbb{R}^{H\times W\times 3}$ is the i-th image, $Y_i \in \{0,1\}^{H\times W}$ is the ground-truth anomaly mask, and $y_i \in \{0,1\}$ is the image-level label. The target testing set $\mathcal{T}$ comprises multiple unseen domains $\mathcal{D}^t_{\text{novel}} = \{(X_i, Y_i, y_i)\}_{i=1}^{N_t}$, characterized by disjoint object categories and novel anomaly types, *i.e.*, $\mathcal{D}_{\text{base}} \cap \mathcal{D}^t_{\text{novel}} = \emptyset$. In zero-shot setting, the model must perform both image anomaly classification and pixel anomaly segmentation on $\mathcal{D}^t_{\text{novel}}$ without any prior exposure to the target domain. In few-shot setting, a minimal support set $\mathcal{P}_c = \{X_i\}_{i=1}^{K}$ consisting of K normal images ($K \in \{1, 2, 4\}$) is provided for each target class to facilitate domain adaptation.

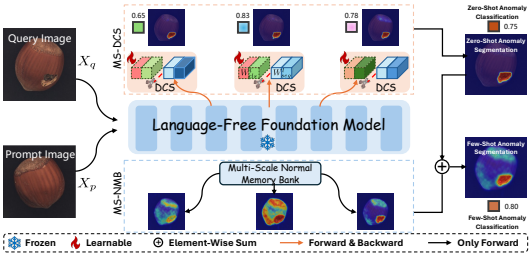


Figure 2: The language-free **UniADet** framework. UniADet decouples global image anomaly classification and local patch segmentation so that their weights are learned independently across hierarchical features. Once trained, it can identify anomalous images and segment anomalous regions in a zero-shot manner. Furthermore, it can be extended to few-shot scenarios using a multi-scale normal memory bank constructed from a few normal reference images, without additional training.

Overview: As illustrated in Fig. 2, the proposed UniADet addresses the challenge of universal AD through Decoupling Classification and Segmentation (DCS, Sec. 3.2), Multi-Scale DCS ((MS-DCS, Sec. 3.3) and a Multi-Scale Normal Memory Bank (MS-NMB, Sec. 3.4). We detail them in the following sections.

3.1 Rethinking VLMs for Anomaly Detection

A VLM-based AD model consists of a visual encoder $\mathcal{F}(\boldsymbol{\theta}_v; \cdot)$ and a text encoder $\mathcal{T}(\boldsymbol{\theta}_t; \cdot)$, where $\boldsymbol{\theta}_v$ and $\boldsymbol{\theta}_t$ are optional, learnable parameters for visual and textual adaptation, respectively. Here, we omit the original parameters for simplicity. Given a query image $X_q \in \mathcal{R}^{H\times W\times 3}$, the visual encoder extracts its representations, yielding a global image token $\boldsymbol{x}_q$ and local patch tokens F_q:

$$\boldsymbol{x}_q, F_q = \mathcal{F}(\boldsymbol{\theta}_v; X_q), \tag{1}$$

where $\boldsymbol{x}_q \in \mathcal{R}^d$, $F_q \in \mathcal{R}^{\frac{H}{P} \times \frac{W}{P} \times d}$, d is the token dimension and P is the patch size.

The text encoder then generates a two-class weight matrix, $W \in \mathcal{R}^{d \times 2}$, using either fixed textual prompts T_f or learnable prompt embeddings T_l, encoding the semantics of normal and anomalous concepts:

$$W = \mathcal{T}(\boldsymbol{\theta}_t; T_i), i \in \{f, l\}. \tag{2}$$

In conventional VLM-based AD, W is shared across classification and segmentation tasks to derive the image-level anomaly score $\hat{y}_z$ and the pixel-level anomaly map $\hat{Y}_z$ via cosine similarity:

$$\hat{y}_z = \sigma\big(\langle W, \boldsymbol{x}_q \rangle / \tau\big), \quad \hat{Y}_z = \sigma\big(\langle W, F_q \rangle / \tau\big), \tag{3}$$

where $\sigma(\cdot)$ denotes the softmax function, $\langle \cdot \rangle$ denotes cosine similarity, and $\tau > 0$ is a temperature parameter.

As summarized in Tab. 1, recent methods are heavily dependent on text encoder, which limits the scalability, making them applicable only to VLMs and unable to be extended to other underlying models, such as DINOv3. Furthermore, these existing approaches tend to aggregate multiple components to achieve competitive performance, which often increases model complexity. In contrast, we explore the feasibility of achieving a simple, general, and scalable AD framework by significantly reducing these components.

Table 1: Component comparison of state-of-the-art Zero-Shot (ZS) and Few-Shot (FS) AD methods. Here, $\mathcal{F}$ and $\mathcal{T}$ denote the visual and textual encoders, respectively. $\boldsymbol{\theta}_v$ and $\boldsymbol{\theta}_t$ represent learnable parameters within their corresponding adapters, while T_f and T_l refer to fixed and learnable textual prompts. ✓ indicates used or satisfied, ✗ indicates not used or satisfied.

Methods	$\mathcal{F}(\cdot)$	$\boldsymbol{\theta}_v$	$\mathcal{T}(\cdot)$	$\boldsymbol{\theta}_t$	T_f	T_l	ZS	FS
WinCLIP [22]	✓	✗	✓	✗	✓	✗	✓	✓
AnomalyCLIP [55]	✓	✗	✓	✓	✗	✓	✓	✗
AdaCLIP [4]	✓	✓	✓	✓	✓	✗	✓	✗
InCtrl [57]	✓	✓	✓	✓	✓	✗	✗	✓
Bayes-PFL [35]	✓	✓	✓	✗	✓	✓	✓	✗
AdaptCLIP [15]	✓	✓	✓	✓	✓	✓	✓	✓
UniADet (ours)	✓	✗	✗	✗	✗	✗	✓	✓

Since W essentially functions as learnable weight parameters in the visual latent space, we hypothesize that the textual encoder is an auxiliary and perhaps redundant intermediary. By directly optimizing W as learnable parameters, we can eliminate the entire textual encoder, then Eq. 2 reduces to a directly parameterized weight matrix:

$$W = [\boldsymbol{w}_n, \boldsymbol{w}_a], \tag{4}$$

where $\boldsymbol{w}_n$ and $\boldsymbol{w}_a$ are optimized directly against the fixed visual features. Moreover, given the strong generalization of modern foundation models, the learnable visual adaptation $\boldsymbol{\theta}_v$ may be superfluous. More importantly, since the W are optimized directly in the visual latent space, it inherently serve as an adaptive probe that aligns with the fixed feature distribution. Thus, Eq. 1 simplifies to a fixed feature extraction process:

$$\boldsymbol{x}_q, F_q = \mathcal{F}(X_q). \tag{5}$$

Substituting Eqs. 4 and 5 into Eq. 3, we implement a simple and general zero-shot AD framework that only needs to learn weights W based on features extracted from foundation models by removing most components (*i.e.*, $\mathcal{T}(\cdot)$, $\boldsymbol{\theta}_v$, $\boldsymbol{\theta}_t$ and T_f) as shown in Tab. 1.

3.2 Decoupling Classification and Segmentation

Several state-of-the-art zero-shot AD methods [55, 4, 35, 15] share a weight matrix W between image-level classification and pixel-level segmentation, as in Eq. 3. However, we argue that this shared-parameter paradigm is suboptimal due to the fundamental distribution differences between global image-level and local patch-level manifolds. As empirically demonstrated by the t-SNE visualizations in Fig. 3(a), the optimal decision hyperplanes for global tokens $\boldsymbol{x}_q$ and local tokens F_q diverge significantly. Crucially, this manifold misalignment is a persistent characteristic observed across both vision-language models (*e.g.*, CLIP) and self-supervised vision-only models (*e.g.*, DINOv2 and DINOv3; see `Appendix` for visualizations).

To mitigate the learning conflict arising from these disparate feature distributions, we propose to explicitly decouple the shared weight W into two task-specific weights, W_{cls} for global classification and W_{seg} for local segmentation. Consequently, the image and pixel anomaly predictions in Eq. 3 are reformulated as follows:

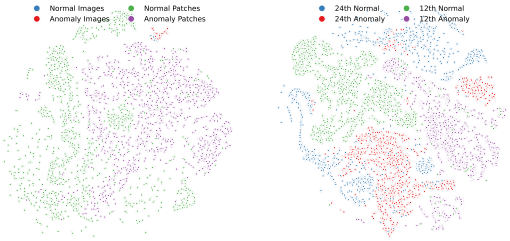
$$\hat{y}_z = \sigma\big(\langle W_{cls}, \boldsymbol{x}_q \rangle / \tau\big), \quad \hat{Y}_z = \sigma\big(\langle W_{seg}, F_q \rangle / \tau\big). \tag{6}$$

This decoupling strategy offers two primary advantages. First, it fundamentally resolves the learning conflict caused by the distinct distributions of global and local representations, allowing each weight to converge to its respective optimal hyperplane. Second, the proposed decoupling paradigm achieves superior performance through a highly parameter-efficient design, bypassing the need for complex adapters or convoluted prompt strategies prevalent in current state-of-the-art methods.

3.3 Multi-Scale Decoupling Classification and Segmentation

For dense visual perception tasks, such as object detection and semantic segmentation, leveraging multi-scale features from a foundation model is a standard paradigm. Most zero/few-shot AD methods follow this approach by aggregating hierarchical features, yet they typically employ a shared weight matrix across these distinct scales.

However, features from different blocks inherently encode varying levels of semantic abstraction, causing them to reside on disparate manifolds, as illustrated in Fig. 3(b). To effectively mitigate the learning conflict arising from these hierarchical distribution shifts, we extend the decoupling principle across all selected blocks. This yields multi-scale decoupling. Let $\boldsymbol{x}_q^l$ and F_q^l denote the global image tokens and local patch tokens extracted from the l-th block, respectively, with W_{cls}^l and W_{seg}^l representing their corresponding block-specific weights. Consequently, the decoupled formulation in Eq. 6 is generalized for each block l as follows:



(a) Global vs. Local (b) Shallow vs. Deep

Figure 3: t-SNE visualization of feature distributions on the MVTec Test Set (Hazelnuts) using CLIP ViT-L/14@336px. (a) A substantial distribution shift is observed between global image tokens ($\boldsymbol{x}^q$) and local patch tokens (F^q) extracted from the 24th block, indicating that representations at different spatial scales reside on distinct manifolds. (b) The manifold of local tokens across layers exhibits significant distribution divergence across different scale (12th vs. 24th blocks), where the clusters of normal and anomalous features at different scales do not overlap within the latent space.

$$\hat{y}_z^l = \sigma\big(\langle W_{cls}^l, \boldsymbol{x}_q^l \rangle / \tau\big), \quad \hat{Y}_z^l = \sigma\big(\langle W_{seg}^l, F_q^l \rangle / \tau\big). \tag{7}$$

This multi-scale decoupling ensures that the anomaly classification and segmentation heads at each hierarchical level can converge to their respective optimal decision boundaries, free from the interference of heterogeneous feature manifolds. Finally, the zero-shot image- and pixel-level predictions ($\hat{y}_z$ and $\hat{Y}_z$) are obtained by averaging the corresponding multi-scale predictions $\{\hat{y}_z^l\}$ and $\{\hat{Y}_z^l\}$ in Eq. 7, respectively.

3.4 Few-Shot UniADet via a Multi-Scale Normal Memory Bank

Although multi-scale decoupled weights $\{W_{seg}^l\}$ and $\{W_{cls}^l\}$ excel at detecting generic anomalies (*e.g.*, scratches or cracks), their efficacy can be constrained when anomalies are defined by subtle deviations from a specific normal reference (*e.g.*, a missing component). To address this, we draw inspiration from the memory-based paradigms in PatchCore [38] and WinCLIP [22], extending our framework to the few-shot setting via a Multi-Scale Normal Memory Bank.

For a set of K normal reference images $X_p \in \mathcal{R}^{K\times H\times W\times 3}$, we first construct a scale-specific memory bank, $M^l = \{\boldsymbol{m}_k^l | \boldsymbol{m}_k^l \in \mathbb{R}^d\}_{k=1}^N$, by collecting N ($K\frac{HW}{P^2}$) d-dimensional local patch tokens $F_p^l \in \mathcal{R}^{K\times \frac{H}{P}\times \frac{W}{P}\times d}$ from the l-th block. Given a query image X_q, we extract its corresponding patch tokens $F_q^l \in \mathbb{R}^{\frac{H}{P}\times \frac{W}{P}\times d}$. The few-shot anomaly score at spatial location (i, j) is determined by the cosine distance to its nearest neighbor in M^l:

$$\hat{Y}_f^l(i,j) = \min_{\boldsymbol{m}_k^l \in M^l} \big(1 - \langle \boldsymbol{F}_q^l(i,j), \boldsymbol{m}_k^l \rangle\big). \tag{8}$$

This process is performed across all selected blocks L. We then compute the multi-scale few-shot anomaly map $\hat{Y}_f$ through simple averaging:

$$\hat{Y}_f = \frac{1}{|L|} \sum \hat{Y}_f^l. \tag{9}$$

To leverage both general anomaly knowledge and reference-specific information, we fuse the zero-shot and few-shot predictions. The final anomaly map is updated as:

$$\hat{Y}_f \leftarrow (1-\lambda_f)\hat{Y}_z + \lambda_f \hat{Y}_f, \tag{10}$$

where $\lambda_f \in [0,1]$ is a balancing coefficient. Similarly, the image-level few-shot anomaly score $\hat{y}_f$ is obtained by fusing $\hat{y}_z$ with the maximum value of $\hat{Y}_f$:

$$\hat{y}_f \leftarrow (1-\lambda_p)\hat{y}_z + \lambda_p \max_{i,j}(\hat{Y}_f(i,j)). \tag{11}$$

3.5 Training

We optimize $\{W^l_{cls}\}$ using cross-entropy loss for global image anomaly classification, and $\{W^l_{seg}\}$ using Focal and Dice losses for local patch anomaly segmentation following AnomalyCLIP [55]. In addition, multi-scale context plays a crucial role in visual anomaly detection because anomalies are essentially defined by their inconsistency with the context. We introduce a Class-Aware image Augmentation (CAA) strategy to improve robustness, which enriches contextual representations through Grid Mosaic and Grid Cropping. Unlike standard augmentations, CAA employs class-specific sampling to ensure that augmented images consistently preserve the original supervisory signals, preventing label noise during the scale-diversification process. Further details regarding the implementation of CAA and the specific training configurations are provided in `Appendix`.

4 Experiments

4.1 Experimental Setup

Datasets: We comprehensively evaluate UniADet across two distinct domains covering industrial inspection and medical diagnosis. In industrial domain, we evaluate six benchmarks with zero-shot and few-shot: MVTec [2], VisA [58], BTAD [34], DTD [1], KSDD [44], and large-scale Real-IAD [47]. In medical domain, we evaluate zero-shot on brain tumor detection (HeadCT, BrainMRI [25], Br35H [19]), skin lesion segmentation (ISIC [6]), gastrointestinal polyp segmentation (ClinicDB [3], ColonDB [45], Kvasir [23], Endo [46]). Detailed statistical information for these datasets can be found in `Appendix`.

Evaluation Metrics: Following previous works, we use AUROC and AUPR for image anomaly classification and pixel anomaly segmentation in our main paper. Here, we emphasize that pixel AUPR is more suitable for anomaly segmentation, where normal and anomalous pixels are highly imbalanced [58]. In `Appendix`, we also provide detailed results using all metrics, including image and pixel AUROC, $F1_{max}$, AUPR, and pixel AUPRO.

Training and Testing Protocol: Following recent works [36, 35], we train UniADet using the testing data from VisA and evaluate zero-/few-shot performance on other datasets. For the VisA evaluation, we use MVTec's test data to train UniADet.

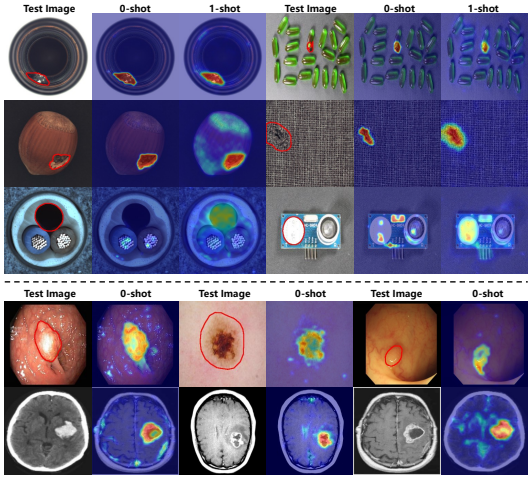


Figure 4: Qualitative comparisons of UniADet using foundation DINOv3 (ViT-L/16) in industrial and medical domains. Best viewed in color and zoom.

4.2 Comparisons with Zero-Shot Methods

Tabs. 2 and 3 compare UniADet with SOTA zero-shot methods on two diverse domains, industrial and medical.

Zero-Shot Generalization in Industrial Domain: As reported in Tab. 2, zero-shot results in industrial domain demonstrate the effectiveness and advantages of our language-free framework, highlighting three significant observations: First, our language-free UniADet substantially outperforms AdaptCLIP [15] (91.6% vs. 89.6%

Table 2: Comparisons with SOTA zero-shot AD methods on industrial datasets. ✓ indicates language-free and ✗ means language-dependent. The best and second-best mean results are highlighted accordingly. †, ‡ and ⋆ indicate CLIP ViT-L/14@336px, DINOv2-R ViT-L/16 and DINOv3 ViT-L/16, respectively, the same below.

Metric	Dataset	WinCLIP[22] ✗	AdaCLIP[4] ✗	AnomalyCLIP[55] ✗	Bayes-PFL[35] ✗	AdaptCLIP[15] ✗	**UniADet†** ✓	**UniADet‡** ✓	**UniADet⋆** ✓
Image-Level (AUROC, AUPR)	MVTec	90.4 , 95.6	90.7 , 95.6	91.6 , 96.4	92.3 , 96.7	93.5 , 96.7	92.4 , 96.1	93.5 , 97.2	94.0 , 97.1
	VisA	75.5 , 78.7	81.7 , 84.3	82.0 , 85.3	87.0 , 89.2	84.8 , 87.6	88.0 , 89.1	91.3 , 91.0	91.9 , 92.8
	BTAD	68.2 , 70.9	89.9 , 95.5	88.3 , 88.2	93.2 , 96.5	91.0 , 92.2	96.4 , 97.8	94.4 , 96.3	94.7 , 97.0
	DTD	95.1 , 97.7	92.7 , 96.4	93.9 , 97.2	95.1 , 98.4	96.0 , 98.4	98.1 , 99.1	94.3 , 98.2	97.1 , 99.0
	KSDD	92.9 , 84.9	96.6 , 89.3	97.8 , 94.2	92.6 , 65.5	98.1 , 95.7	96.3 , 83.4	91.6 , 59.5	91.0 , 64.5
	Real-IAD	67.0 , 62.9	73.3 , 70.2	69.5 , 64.3	70.0 , 65.8	74.2 , 70.8	78.6 , 76.7	82.5 , 80.3	81.2 , 79.3
	Mean	81.5 , 81.8	87.5 , 88.6	87.2 , 87.6	88.4 , 85.3	89.6 , **90.2**	**91.6** , **90.4**	91.3 , 87.1	**91.7** , 88.3
Pixel-Level (AUROC, AUPR)	MVTec	82.3 , 18.2	88.3 , 39.1	91.1 , 34.5	91.8 , 48.3	90.9 , 38.3	91.8 , 42.8	92.4 , 50.9	93.2 , 52.7
	VisA	73.2 , 5.4	95.7 , 31.0	95.5 , 21.3	95.6 , 29.8	95.7 , 26.1	95.8 , 28.0	95.9 , 32.7	96.3 , 32.5
	BTAD	72.7 , 12.9	91.6 , 42.9	94.2 , 45.5	93.9 , 47.1	93.8 , 41.8	95.7 , 42.8	97.7 , 60.6	97.8 , 58.9
	DTD	79.5 , 9.8	98.3 , 75.2	97.9 , 62.6	97.8 , 69.9	97.7 , 68.7	98.4 , 64.3	98.3 , 71.9	99.1 , 75.9
	KSDD	93.0 , 7.1	97.6 , 48.2	98.1 , 51.9	97.7 , 18.7	98.1 , 58.1	98.1 , 31.8	98.4 , 15.8	99.0 , 22.6
	Real-IAD	84.5 , 3.3	96.1 , 30.5	95.1 , 26.7	96.5 , 27.6	94.9 , 28.2	96.6 , 33.6	97.7 , 43.1	97.5 , 41.6
	Mean	80.9 , 9.5	94.6 , 44.5	95.3 , 40.4	95.5 , 40.2	95.2 , 43.5	96.1 , 40.6	**96.7**, **45.8**	**97.2** , **47.4**

Table 3: Comparisons with SOTA zero-shot AD methods on medical datasets.

Metric	Dataset	WinCLIP[22] ✗	AdaCLIP[4] ✗	AnomalyCLIP[55] ✗	Bayes-PFL[35] ✗	AdaptCLIP[15] ✗	**UniADet†** ✓	**UniADet‡** ✓	**UniADet⋆** ✓
Image-Level (AUROC, AUPR)	HeadCT	83.7 , 81.6	93.4 , 92.2	93.0 , 92.6	96.5 , 95.5	92.6 , 91.9	97.0 , 96.7	95.2 , 94.9	95.7 , 94.8
	BrainMRI	92.0 , 90.7	94.9 , 94.2	90.2 , 92.4	96.2 , 92.4	90.1 , 92.0	95.7 , 96.4	91.7 , 94.8	91.9 , 95.0
	Br35H	80.5 , 82.2	95.7 , 95.7	94.2 , 94.2	97.8 , 96.2	94.8 , 95.1	97.4 , 97.1	92.5 , 93.9	92.3 , 93.8
	Mean	85.4 , 84.8	94.7 , 94.0	92.5 , 93.1	**96.8** , **94.7**	92.5 , 93.0	**96.7** , **96.7**	93.1 , 94.5	93.3 , 94.5
Pixel-Level (AUROC, AUPR)	ISIC	83.3 , 62.4	85.4 , 70.6	89.4 , 76.2	92.2 , 84.6	91.0 , 80.9	92.6 , 84.2	95.0 , 88.5	95.9 , 90.6
	ColonDB	64.8 , 14.3	79.3 , 26.2	81.9 , 31.7	82.1 , 31.9	86.5 , 52.0	84.0 , 36.2	86.0 , 35.5	86.7 , 33.3
	ClinicDB	70.7 , 19.4	84.3 , 36.0	82.9 , 38.2	89.6 , 53.2	83.8 , 40.2	90.1 , 55.8	91.3 , 59.3	91.5 , 57.5
	Endo	68.2 , 23.8	84.0 , 44.8	84.2 , 46.6	89.2 , 58.6	82.1 , 45.3	90.7 , 62.5	92.6 , 67.3	92.8 , 68.5
	Kvasir	69.8 , 27.5	79.4 , 43.8	79.1 , 39.8	85.4 , 54.2	86.5 , 52.0	87.5 , 57.5	91.1 , 65.2	91.7 , 65.9
	Mean	71.4 , 29.5	82.5 , 44.3	83.5 , 46.5	87.7 , 56.5	86.0 , 54.1	89.0 , 59.2	**91.2** , **63.2**	**91.7** , **63.2**

I-AUROC) in image anomaly classification using the same CLIP model. This validates our core hypothesis: *complex language prompts, textual encoders, and elaborate adapters are unnecessary*. Second, *our language-free framework enables a seamless switch to advanced vision models (*e.g.*, DINOv2/v3)*, achieving a 5.6%-7.2% P-AUPR lead over the SOTA Bayes-PFL [35]. This confirms that self-supervised representations are more effective for pixel-level localization than those derived from multi-modal CLIP alignment. Third, our UniADet‡ exceeds AdaptCLIP on the large-scale Real-IAD [47] benchmark by 8.3% I-AUROC and 14.9% P-AUPR. This substantial performance margin further validates the *stability and scalability* of our language-free framework when addressing complex and real-world scenarios.

Cross-Domain Zero-Shot Generalization to Medical Data: To further validate the generalization capability of our framework, we evaluate UniADet on medical datasets after training it exclusively on an industrial auxiliary dataset (*e.g.*, VisA), representing a challenging cross-domain generalization task. As reported in Tab. 3, UniADet achieves highly competitive performance across eight medical benchmarks, establishing a new SOTA in most scenarios despite its embarrassingly simple and efficient design (0.02M parameter). Notably, in pixel anomaly segmentation, UniADet utilizing the DINOv3 backbone achieves 91.7% P-AUROC and 63.2% P-AUPR, exceeding SOTA Bayes-PFL [35] by a significant margin. These results decisively demonstrate the robust and superior cross-domain transferability of our language-free framework.

Complexity and Efficiency: We evaluate model complexity based on the total number of parameters (frozen + learnable) and assess efficiency using the average inference time per image, as summarized in Tab. 4. All benchmarks are conducted on a single H20 GPU with a batch size of 1. Notably, UniADet drastically reduces the number of learnable parameters by approximately two and three orders of magnitude compared to the SOTA AdaptCLIP [15] and Bayes-PFL [35], respectively. In terms of efficiency, UniADet achieves a superior average inference time of 15.7 ms per image using CLIP ViT-L/14@336px model, outperforming exist-

Table 4: Complexity and efficiency comparisons.

Shots	Methods	Models	Input Size	# Params (M)	Inf.Time (ms)
0	AdaCLIP [4]	CLIP ViT-L/14@336px	518×518	428.8 + 1.1e+1	107.4
	AnomalyCLIP [55]	CLIP ViT-L/14@336px	518×518	427.9 + 5.6e+0	70.7
	Bayes-PFL [35]	CLIP ViT-L/14@336px	518×518	427.9 + 2.7e+1	154.9
	AdaptCLIP-Zero	CLIP ViT-L/14@336px	518×518	427.9 + 6.0e-1	57.5
	UniADet†	CLIP ViT-L/14@336px	518×518	342.9 + **1.5e-2**	**15.7**
	UniADet⋆	DINOv3 ViT-L/16	512×512	303.2 + 2.0e-2	41.9
1	InCtrl [57]	CLIP ViT-B-16+240	240×240	208.4 + 3.0e-1	59.0
	AnomalyCLIP+ [55]	CLIP ViT-L/14@336px	518×518	427.9 + 5.6e+0	76.2
	AdaptCLIP	CLIP ViT-L/14@336px	518×518	427.9 + 1.8e+0	58.7
	UniADet†	CLIP ViT-L/14@336px	518×518	342.9 + **1.5e-2**	**22.4**
	UniADet⋆	DINOv3 ViT-L/16	512×512	303.2 + 2.0e-2	48.4

ing methods by a significant margin. Furthermore, our framework exhibits remarkable scalability: extending from a zero-shot to a few-shot setting introduces negligible computational overhead, requiring no additional learnable parameters and incurring minimal latency increase.

Qualitative Results: Fig. 4 shows some visualizations from industrial and medical testing images using UniADet with a foundation DINOv3 (ViT-L/16). Generally, one-shot setting improves the performance over zero-shot setting, especially for reference-dependent anomalies that are defined by their deviation from a specific normal reference (shown in the third row).

4.3 Comparisons with Few- and Full-Shot Methods

Tab. 5 compares UniADet under few-shot settings against SOTA few-shot and full-shot AD models with "I-AUROC / P-AUPR" metrics. For almost all methods, the transition from zero-shot to few-shot, yields substantial performance gains. This underscores that normal reference images are pivotal to industrial AD, as many anomalies are fundamentally defined by their structural deviation from the norm.

Our framework demonstrates several key advantages: *SOTA performance with minimal resources.* Our 4-shot UniADet achieves the best overall performance in Tab. 5 by utilizing only a single DINOv2 backbone. It significantly outperforms VLM-based methods, highlighting the superior efficiency of our decoupling strategy and training-free multi-scale memory bank. Crucially, our approach also consistently outperforms SOTA FOUNDAD [52] built upon DINOv3. *Competitive against heavy ensembles.* UniADet exceeds UniVAD [17], which uses a five-model ensemble. This demonstrates that a well-decoupled single foundation model can be more effective and efficient than heavy multi-model integration. *Challenging full-shot paradigms.* Using only 4 normal reference images and requiring no target-specific training, UniADet significantly surpasses established full-shot methods like UniAD [50]. Furthermore, our results are highly competitive with the current full-shot SOTA, Dinomaly [18]. Remarkably, on the more challenging and diverse Real-IAD benchmark, our 4-shot results actually *exceed* the full-shot Dinomaly (90.3% vs. 89.3% I-AUROC). These findings demonstrate that UniADet, while remaining training-free and parameter-efficient, unlocks the immense potential of vision foundation models for practical, open-world few-shot anomaly detection.

Table 5: Comparisons with state-of-the-art few-/full-shot AD methods. Gray indicates models trained on all available normal training images. Additional comparisons are provided in `Appendix`.

Methods	Shots	MVTec	VisA	Real-IAD
UniADet‡	1	97.6 / 63.1	95.2 / 42.1	88.7 / 48.4
	4	**98.7 / 65.4**	**96.9 / 45.2**	**90.3 / 48.5**
UniVAD [17]	1	97.8 / 55.6	93.5 / 42.8	85.1 / 37.6
FOUNDAD [52]	1	96.1 / -	92.6 / -	- / -
	4	97.1 / -	94.4 / -	- / -
AdaptCLIP[15]	1	94.5 / 53.7	90.5 / 38.9	81.8 / 36.6
	4	96.6 / 57.2	93.1 / 41.8	83.9 / 39.1
MetaUAS [13]	1	90.7 / 59.3	81.2 / 42.7	80.0 / 36.6
InCtrl [57]	1	91.3 / 47.8	83.2 / 17.7	76.6 / 19.1
	4	93.1 / 50.9	87.8 / 19.2	81.8 / 21.0
APRIL-GAN [5]	1	92.0 / 51.8	91.2 / 30.9	- / -
	4	92.8 / 54.5	92.6 / 32.2	- / -
WinCLIP [22]	1	93.6 / 38.3	80.0 / 15.8	74.7 / 13.9
	4	95.3 / 41.2	84.3 / 18.1	77.0 / 15.4
Dinomaly [18]	full	99.6 / 69.3	98.7 / 53.2	89.3 / 42.8
UniAD [50]	full	96.5 / 44.7	90.8 / 33.6	83.0 / 21.1

4.4 Ablation Study

To demonstrate the effectiveness of UniADet, we conduct extensive ablation studies (Tab. 6) on its core components: Decoupling Classification and Segmentation (DCS), Multi-Scale DCS (MS-DCS), and Class-Aware Augmentation (CAA). Furthermore, to evaluate the versatility and plug-and-play capability of our proposed modules, we integrate DCS and MS-DCS into the language-dependent AnomalyCLIP. By comparing this augmented baseline with our approach, we decouple the intrinsic gains of the language-free paradigm from purely structural improvements.

DCS: In Tab. 6(a), we first establish a simple baseline (Line a) by removing the CLIP text encoder and utilizing a shared weight for both global anomaly classification and local anomaly segmentation across all feature levels. This baseline already outperforms AnomalyCLIP [55] at the pixel level. Crucially, when we decouple anomaly classification and segmentation by learning two independent weights (Line b vs. a), both image- and pixel-level performance improve significantly. This empirically verifies that the decoupling strategy effectively mitigates the inherent conflict between the two tasks, which arises from their distinct feature manifolds.

MS-DCS: We further extend the decoupling concept to encompass all selected scales, where each scale utilizes its own independent weights (Line c in Tab. 6(a)). This configuration leads to further

Table 6: **Ablation studies using CLIP ViT-L/14@336px**. Performance is measured using I-AUROC and P-AUPR metrics. Default settings are marked in blue.

(a) Effect of different components.

ID	DCS	MS-DCS	CAA	Shot	MVTec	VisA
a	✗	✗	✗	0	85.4 / 36.4	77.9 / 26.1
b	✓	✗	✗	0	91.8 / 38.3	85.9 / 27.2
c	✓	✓	✗	0	92.2 / 40.7	86.0 / 27.6
d	✓	✓	✓	0	92.4 / 42.8	88.0 / 28.0
e	✓	✓	random	0	91.3 / 41.5	87.5 / 26.6

(b) Effect of the selected feature layers.

ID	Blocks	Shot	MVTec	VisA
a	{24}	0	89.9 / 35.4	85.3 / 23.1
b	{21, 24}	0	92.1 / 36.7	87.3 / 24.2
c	{18, 21, 24}	0	92.0 / 38.9	87.7 / 26.1
d	{15, 18, 21, 24}	0	92.2 / 41.1	87.6 / 26.9
e	{12, 15, 18, 21, 24}	0	92.4 / 42.8	88.0 / 28.0

(c) Models trained on MVTec and tested on VisA. LP means the number of learnable parameters (Million).

ID	DCS	MS-DCS	CAA	Blocks	**AnomalyCLIP** Style		**UniADet**† Style		Language-Free
					Performance ↑	# LP ↓	Performance ↑	# LP ↓	Gain
a	✗	✗	✗	{24}	82.0 / 21.3	5.6	82.4 / 21.1	0.0015	**↑0.4 / ↓0.2**
b	✓	✗	✗	{24}	83.3 / 22.1 ↑1.3/↑0.8	11.2	85.3 / 23.1 ↑2.9/↑2.0	0.0030	**↑2.0 / ↑1.0**
c	✗	✗	✗	{12, 15, 18, 21, 24}*	72.3 / 20.6 ↓9.7/↓0.7	5.6	77.9 / 26.1 ↓4.5/↑5.0	0.0015	**↑5.6 / ↑5.5**
d	✓	✓	✗	{12, 15, 18, 21, 24}*	83.9 / 25.9 ↑1.9/↑4.6	56.0	86.0 / 27.6 ↑3.6/↑6.5	0.0154	**↑2.1 / ↑1.7**

* As clarified in AnomalyCLIP's official repository (Issue #18), the initial results were based on a single layer (24th) despite the multi-scale design. Our results reflect its performance when correctly extended to multi-block, where it suffers from scale-wise distribution shifts.

performance gains, conclusively confirming that the full decoupling mechanism effectively mitigates task conflicts arising from disparate feature manifolds both within a layer (global vs. local) and across different blocks. Ultimately, this multi-scale decoupling strategy successfully unlocks the full potential of the foundation model for universal anomaly detection.

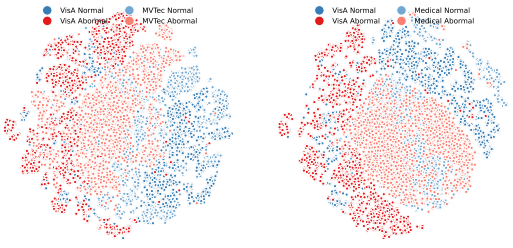
Language-Free: We integrate DCS and MS-DCS into language-dependent AnomalyCLIP and compare the results with UniADet in Tab. 6(c). Our language-free paradigm consistently outperforms its counterpart, AnomalyCLIP, across nearly all settings, validating that neither an explicit language encoder or prompts are necessary for visual AD. Both DCS and MS-DCS provide consistent gains even when integrated into the heavy AnomalyCLIP, proving their plug-and-play effectiveness. While a naive multi-block extension (Line c) triggers a performance drop due to scale-wise distributional shifts, DCS and MS-DCS exhibit superior robustness (Line d). Notably, UniADet achieves these results with orders of magnitude fewer parameters (0.015M vs. 56.0M), highlighting its efficiency.

Multi-Scale Features: We analyze the influence of integrating hierarchical features from different blocks on overall zero-shot AD performance, as reported in Tab. 6(b). The performance generally improves as multi-scale features are incorporated, validating the fundamental assumption that hierarchical features provide rich, complementary information crucial for comprehensive AD. In particular, fine-grained pixel-level segmentation relies heavily on the low-level details preserved in shallower blocks (*e.g.*, the 12th and 15th blocks).

CAA: The class-aware augmentation yields further performance enhancement by incorporating multi-scale context (Line d in Tab. 6(a)). In contrast, we find that employing a class-agnostic augmentation strategy (*i.e.*, random augmentation in Line e) results in performance degradation relative to CAA. This empirical evidence confirms that maintaining semantic consistency across augmented contexts is vital for robust learning in visual anomaly detection.

4.5 Why Does UniADet Work?

Generalization from Industrial to Medical: The robust generalization of UniADet stems from its minimalist architecture and dual-decoupling strategy (DCS and MS-DCS). We hypothesize that despite the diversity of industrial defects and medical lesions, they share a common distribution characteristic in the visual latent space: they are inherently distinct from normal patterns. To validate this, we employ t-SNE to visualize the distribution of local patch



(a) VisA & MVTec (b) VisA & Medical

Figure 5: t-SNE visualizations from datasets within and across domains using DINOv2R ViT-L/14. More visualizations with various backbones are available in Fig. A1.

tokens across different datasets (VisA vs. MVTec) and cross-domain scenarios (VisA vs. Medical). As illustrated in Fig. 5, a consistent pattern emerges: normal features consistently aggregate into compact clusters, whereas anomalous features remain scattered with clear decision boundaries. This observation explains why UniADet achieves superior zero-shot performance without domain-specific training and highlights the intrinsic advantage of a language-free framework.

Learnable Weights vs. Manual Textual Embeddings: To investigate the intrinsic relationship between our learned task-specific weights (*i.e.*, W^l_{cls} and W^l_{seg}) and two-class textual embeddings (*i.e.*, W_{text} as defined in WinCLIP), we conduct a comparative analysis as follows.

Specifically, W_{text} typically comprises embeddings for "normal" and "abnormal" semantics. We decompose W_{text} into $\boldsymbol{w}_a$ and $\boldsymbol{w}_n$, and define the semantic direction as $\Delta\boldsymbol{w}_{text} = \boldsymbol{w}_a - \boldsymbol{w}_n$. Following the same protocol, we derive the learned directions for anomaly classification and segmentation as $\Delta\boldsymbol{w}^l_{cls}$ and $\Delta\boldsymbol{w}^l_{seg}$, respectively. Finally, we employ cosine similarity to quantify the correlation between these learned directions and the textual prior, as visualized in Fig. 6. The empirical results reveal a significant semantic misalignment. The global classification weights ($\Delta\boldsymbol{w}^l_{cls}$) exhibit a strong positive correlation with $\Delta\boldsymbol{w}_{text}$, peaking at the final layer. This confirms that the global image token in CLIP remains highly aligned with the textual semantic space. In contrast, the local segmentation weights ($\Delta\boldsymbol{w}^l_{seg}$) consistently show a negative correlation with the textual direction. This divergence provides empirical evidence of an inherent task conflict between anomaly classification and segmentation. Our findings suggest that generic textual priors fail to capture the fine-grained structural requirements of local anomaly segmentation, thereby underscoring the necessity and superiority of our dual-decoupling strategy (DCS and MS-DCS) in mitigating task interference and enhancing detection precision.

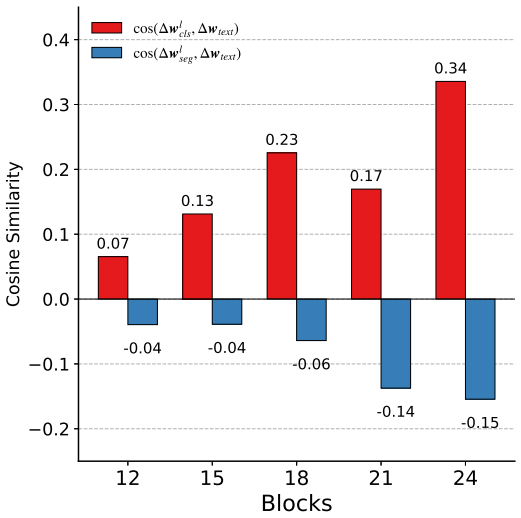


Figure 6: Cosine similarity between our learned decision directions and WinCLIP-derived textual directions.

5 Conclusion

This paper fundamentally challenges and rethinks the prevailing reliance on language-dependent vision anomaly detection. Our core insight reveals that the primary function of a language encoder is merely to derive decision weights for anomaly classification and segmentation. Consequently, we believe that the language encoder may be unnecessary, as these weights can be learned directly and more efficiently through a data-driven process, thus negating the need for laborious prompt engineering, elaborate adaptation modules, and complex training strategies that have been the focus of recent research. This paradigm shift allows us to present UniADet, a remarkably *simple, effective, efficient, general, and language-free* framework for universal anomaly detection. Specifically, this paper introduces an embarrassingly simple method to systematically decouple the learning process: cross-task (classification and segmentation) and cross-scale (hierarchical features) for addressing learning conflicts arising from multi-task and multi-scale distribution shift. By learning independent weights for distinct tasks and individual features, UniADet completely eliminates task-specific and cross-scale interference. Furthermore, our language-free paradigm is inherently more versatile, seamlessly integrating with diverse foundation models, ranging from vision-language CLIP to pure-vision DINOv2 and DINOv3. Extensive experiments across industrial and medical domains confirm that UniADet achieves SOTA performance in both zero-shot and few-shot anomaly detection. Notably, on the large-scale real-world industrial Real-IAD benchmark, our UniADet using only 4 normal images without any target dataset-specific training outperforms full-shot SOTA method for the first time. This empirical evidence reveals the powerful potential of a language-free framework in visual anomaly detection. However, UniADet's language-free nature may limit its reasoning in text-dependent scenarios, and the memory bank's complexity scales with the size of normal set.

References

[1] Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *WACV*, 2023.

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec-AD: A comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 2019.

[3] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43, 2015.

[4] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. AdaCLIP: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *ECCV*, 2024.

[5] Xuhai Chen, Yue Han, and Jiangning Zhang. April-GAN: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. In *CVPRW*, 2023.

[6] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *ISBI*, 2018.

[7] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024.

[8] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: a patch distribution modeling framework for anomaly detection and localization. In *ICPR*, 2021.

[9] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *CVPR*, 2022.

[10] Choubo Ding, Guansong Pang, and Chunhua Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *CVPR*, 2022.

[11] Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. FastRecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *ICCV*, 2023.

[12] Bin-Bin Gao. Learning to detect multi-class anomalies with just one normal image prompt. In *ECCV*, 2024.

[13] Bin-Bin Gao. MetaUAS: Universal anomaly segmentation with one-prompt meta-learning. *NeurIPS*, 2024.

[14] Bin-Bin Gao. Dual-masked and discriminative reconstruction for unified vision anomaly detection. *IEEE Transactions on Image Processing*, 35, 2026.

[15] Bin-Bin Gao, Yue Zhou, Jiangtao Yan, Yuezhi Cai, Weixi Zhang, Meng Wang, Jun Liu, Yong Liu, Lei Wang, and Chengjie Wang. AdaptCLIP: Adapting clip for universal visual anomaly detection. In *AAAI*, 2026.

[16] Tao Gong, Qi Chu, Bin Liu, Wei Zhou, and Nenghai Yu. FE-CLIP: Frequency enhanced clip model for zero-shot anomaly detection and segmentation. In *ICCV*, 2025.

[17] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. UniVAD: A training-free unified model for few-shot visual anomaly detection. In *CVPR*, 2025.

[18] Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *CVPR*, 2025.

[19] Ahmed Hamada. Br35H: Brain tumor detection 2020, 2020.

[20] Jianfang He, Min Cao, Silong Peng, and Qiong Xie. RareCLIP: Rarity-aware online zero-shot industrial anomaly detection. In *ICCV*, 2025.

[21] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *ECCV*, 2022.

[22] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. WinCLIP: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, 2023.

[23] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-SEG: A segmented polyp dataset. In *MMM*, 2019.

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

[25] Driss Lamrani, Bouchaib Cherradi, Oussama El Gannour, Mohammed Amine Bouqentar, and Lhoussain Bahatti. Brain tumor detection using mri images and convolutional neural network. *International Journal of Advanced Computer Science and Applications*, 13(7), 2022.

[26] Jiarui Lei, Xiaobo Hu, Yue Wang, and Dong Liu. PyramidFlow: High-resolution defect contrastive localization using pyramid normalizing flow. In *CVPR*, 2023.

[27] Aodong Li, Chen Qiu, Marius Kloft, Padhraic Smyth, Maja Rudolph, and Stephan Mandt. Zero-shot anomaly detection via batch normalization. In *NeurIPS*, 2023.

[28] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. CutPaste: Self-supervised learning for anomaly detection and localization. In *CVPR*, 2021.

[29] Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. PromptAD: Learning prompts with only normal samples for few-shot anomaly detection. In *CVPR*, 2024.

[30] Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. MuSc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In *ICLR*, 2024.

[31] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024.

[32] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. SimpleNet: A simple network for image anomaly detection and localization. In *CVPR*, 2023.

[33] Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. AACLIP: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *CVPR*, 2025.

[34] 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 *ISIE*, 2021.

[35] Zhen Qu, Xian Tao, Xinyi Gong, ShiChen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *CVPR*, 2025.

[36] Zhen Qu, Xian Tao, Mukesh Prasad, Fei Shen, Zhengtao Zhang, Xinyi Gong, and Guiguang Ding. VCP-CLIP: A visual context prompting model for zero-shot anomaly segmentation. In *ECCV*, 2024.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[38] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, 2022.

[39] Bo-Kai Ruan, Yi-Zeng Fang, Hong-Han Shuai, and Juinn-Dar Huang. Anomaly detection for hybrid butterfly subspecies via probability filtering. *arXiv preprint arXiv:2504.01671*, 2025.

[40] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *WACV*, 2022.

[41] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *CVPR*, 2021.

[42] Shelly Sheynin, Sagie Benaim, and Lior Wolf. A hierarchical transformation-discriminating generative model for few shot anomaly detection. In *ICCV*, 2021.

[43] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *Transactions on Machine Learning Research*, 2026.

[44] Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3):759–776, 2020.

[45] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging*, 35(2), 2015.

[46] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, Aaron Courville, et al. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, 2017.

[47] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-IAD: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *CVPR*, 2024.

[48] Xincheng Yao, Ruoqi Li, Zefeng Qian, Yan Luo, and Chongyang Zhang. Focus the Discrepancy: Intra- and inter-correlation learning for image anomaly detection. In *ICCV*, 2023.

[49] Xincheng Yao, Ruoqi Li, Jing Zhang, Jun Sun, and Chongyang Zhang. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *CVPR*, 2023.

[50] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In *NeurIPS*, 2022.

[51] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM: A discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, 2021.

[52] Guangyao Zhai, Yue Zhou, Xinyan Deng, Lars Heckler, Nassir Navab, and Benjamin Busam. Foundation visual encoders are secretly few-shot anomaly detectors. In *ICLR*, 2026.

[53] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *CVPR*, 2024.

[54] Ying Zhao. OmniAL: A unified cnn framework for unsupervised anomaly localization. In *CVPR*, 2023.

[55] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *ICLR*, 2024.

[56] Jiawen Zhu, Yew-Soon Ong, Chunhua Shen, and Guansong Pang. Fine-grained abnormality prompt learning for zero-shot anomaly detection. In *CVPR*, 2025.

[57] Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *CVPR*, 2024.

[58] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *ECCV*, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This paper fundamentally challenges and rethinks the prevailing reliance on language-dependent vision anomaly detection, and then proposes UniADet, a simple, effective, efficient, general, and language-free framework for universal (zero-shot and few-shot) vision anomaly detection.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of our work are discussed in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: Our work does not include theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Code and Models are available at `https://github.com/gaobb/UniADet`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code and Models are available at `https://github.com/gaobb/UniADet`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We give implementation details in Section A2 and competing methods in Section A4 in the Supplemental Material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In the few-shot setting, performance may vary with the selected normal reference images. We randomly sample K (*e.g.*, 1, 2, or 4) normal images per class and report the mean and standard deviation over three runs with different random seeds. Due to space limitations, we report the mean results in the main paper, and both the mean and standard deviations in the Supplementary Materials.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We measure the complexity and efficiency with the number of parameters and forward inference times. The evaluation is performed on a single H20 GPU. Please refer to detailed comparisons in Tab. 4.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully read the NeurIPS Code of Ethics. Our research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work may benefit industrial quality inspection and medical image analysis, as discussed in Section 1.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original papers for all datasets: industrial (MVTec [2], VisA [58], BTAD [34], DTD [1], KSDD [44], and Real-IAD [47]) and medical (BrainMRI [25], Br35H [19], ISIC [6], ClinicDB [3], ColonDB [45], Kvasir [23], and Endo [46]). All assets are used in accordance with their licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Our work does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our work involves neither crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our work involves neither crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: This research does not involve LLMs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

A1 Class-Aware Augmentation

Multi-scale context plays a crucial role in visual anomaly detection because the essence of anomalies lies in their inconsistency with the context. We propose a Class-Aware Augmentation (CAA) method, including Grid Mosaic and Grid Cropping, to enhance the model's robustness.

Grid Mosaic: Given a candidate image and its corresponding category, we construct an augmented image by randomly sampling $n^2 - 1$ additional images from the training set and arranging them, together with the candidate image, in an $n \times n$ grid. To ensure the image-level ratio of normal and anomalous images remains unbiased, we apply distinct sampling strategies based on the current image's label (*i.e.*, normal or anomaly). If the candidate image is normal, we randomly sample normal images of the same category as mosaic components. If the candidate image is an anomaly, we randomly sample images of the same category, which may include a mix of both normal and anomalous images. This conditional sampling approach is crucial for maintaining the ratio of normal to anomalous images and focusing the learning process on the true contextual discrepancies that define an anomaly.

Grid Cropping: To obtain representations at different scales and effectively simulate resolution reduction from combining multiple images, we introduce a Grid Cropping Augmentation. Specifically, the current image is first randomly partitioned into an $n \times n$ grid. Subsequently, we randomly select one of these resulting image patches to serve as the augmented input. To ensure the ratio of normal to anomalous images is preserved during this process, we employ different strategies for normal and anomalous images. For a normal image, we select a patch at random. For an anomalous image, we strictly select only those patches that contain an anomalous region. This focused sampling strategy ensures that the model receives reliable supervisory signals for localization while effectively learning features across various scales and resolutions.

Table A1: Key statistics of industrial and medical datasets with different attributes. ✓ means satisfied and ✗ means not satisfied.

Domain	Dataset	Modality	Category	# Classes	Pose-Agnostic	Annotations		Train	Test	
						Image	Pixel	# Normal	# Normal	# Anomaly
Industrial	MVTec [2]	Photography	Obj & Texture	15	✗	✓	✓	3,629	467	1,258
	VisA [58]	Photography	Obj	12	✗	✓	✓	8,659	962	1,200
	BTAD [34]	Photography	Obj & Texture	3	✗	✓	✓	1,799	451	290
	DTD [1]	Photography	Texture	12	✗	✓	✓	1,200	357	947
	KSDD [44]	Photography	Texture	1	✗	✓	✓	857	286	54
	Real-IAD [47]	Photography	Obj	30	✗	✓	✓	36,465	63,256	51,329
Medical	HeadCT [41]	CT	Head	1	✓	✓	✗	0	100	100
	BrainMRI [25]	MRI	Brain	1	✓	✓	✗	0	155	98
	Br35H [19]	MRI	Brain Tumor	1	✓	✓	✗	0	1,500	1,500
	ISIC [6]	Endoscopy	Skin Lesion	1	✓	✗	✓	0	0	379
	ColonDB [45]	Endoscopy	Colorectal Cancer	1	✓	✗	✓	0	0	380
	ClinicDB [3]	Endoscopy	Colorectal Cancer	1	✓	✗	✓	0	0	612
	Endo [46]	Endoscopy	Gastrointestinal Tract	1	✓	✗	✓	0	0	200
	Kvasir [23]	Endoscopy	Gastrointestinal Tract	1	✓	✗	✓	0	0	1,000

A2 Implementation Details

Diverse Datasets: To validate the effectiveness of our **UniADet**, we conduct comprehensive experiments on 14 public anomaly detection benchmarks covering two domains, industrial and medical. Following previous works, we use the industrial VisA [58] as an auxiliary training set to train our **UniADet**, and then evaluate its zero-shot or few-shot generalization on 5 industrial benchmarks, MVTec [2], BTAD [34], DTD [1], KSDD [44], and Real-IAD [47], and 8 medical benchmarks, HeadCT [41], BrainMRI [25], Br35H [19], ISIC [6], ColonDB [45], ClinicDB [3], Kvasir [23], and Endo [46]. For the evaluation of VisA [58], we use the industrial MVTec [2] as the auxiliary training set. The statistical information of all these datasets is reported in Tab. A1.

It should be noted that Real-IAD [47] is a large-scale anomaly detection dataset, consisting of diverse categories (30 categories) and large-scale images (150k) in industrial domain. In medical domain, either image-level annotations or pixel-level annotations are only provided. Therefore, we only report image anomaly classification performance on HeadCT [41], BrainMRI [25] and Br35H [19], and provide pixel-level anomaly segmentation performance on ISIC [6], ColonDB [45], ClinicDB [3], Kvasir [23], and Endo [46]. Furthermore, we report only zero-shot results on the medical benchmarks because their training splits do not provide normal reference images.

Training Details: We evaluate a variety of foundation models, such as visual-language models, CLIP (ViT-L/14@336) [37], and self-supervised vision-only models, DINOv2R (ViT-L/14) [7] and DINOv3 (ViT-L/16) [43], due to the generality of our method. We extract global image tokens and local patch tokens from multiple blocks $\{12, 15, 18, 21, 24\}$. When learning the decoupled weights, we apply CAA to training images with a probability of 0.5 and randomly sample grid sizes from the set $\{2 \times 2, 3 \times 3\}$ with equal probability. We set λ_p to 0.5

and λ_f to 0.8 by default in few-shot setting. During training and testing stages, we resize images to 518×518 and 512×512 for ViT-L/14-based and ViT-L/16-based foundational models, respectively. We train the model for 15 epochs with a learning rate of 0.001, and report the experimental results based on the 15th model. All experiments are conducted using PyTorch with a single NVIDIA H20 GPU.

A3 Comparison between Transductive and Inductive Zero-Shot Settings

MuSc [30] considers transductive zero-shot AD. It is given the class identity and the collection of unlabeled test images from that class, and its Mutual Scoring lets these test images score one another. Its effectiveness relies on dataset-level information and on the assumptions that normal patterns dominate the test set and that anomalous patterns do not recur sufficiently often to appear normal through mutual matching. Its time and memory costs also increase with the size of the test set and may become prohibitive or even cause out-of-memory (OOM) failures on large datasets (e.g., Br35H). By contrast, UniADet performs inductive zero-shot AD: each query is processed independently, without access to other test images. Our cross-domain medical evaluation using the same CLIP ViT-L/14@336px further illustrates this distinction:

Table A2: Performance, memory consumption, and inference time comparison between MuSc and UniADet on medical datasets.

Level	Dataset	MuSc [30]			UniADet		
		Metric (AUROC / AUPR)	Mem.	Time	Metric (AUROC / AUPR)	Mem.	Time
Image	HeadCT	64.9 / 57.9	18,805 MB	873.0 ms	97.0 / 96.7	1,392 MB	15.7 ms
	BrainMRI	81.8 / 82.6	56,421 MB	713.8 ms	95.7 / 96.4	1,392 MB	15.7 ms
	Br35H	OOM	OOM	OOM	97.4 / 97.1	1,392 MB	15.7 ms
Pixel	ISIC	78.8 / 49.4	18,077 MB	1,158.2 ms	92.6 / 84.2	1,392 MB	15.7 ms
	ColonDB	78.3 / 20.0	18,123 MB	1,119.4 ms	84.0 / 36.2	1,392 MB	15.7 ms
	ClinicDB	78.8 / 27.8	24,755 MB	1,852.2 ms	90.1 / 55.8	1,392 MB	15.7 ms
	Endo	81.7 / 36.6	12,967 MB	697.5 ms	90.7 / 62.5	1,392 MB	15.7 ms
	Kvasir	76.0 / 32.1	35,807 MB	2,325.7 ms	87.5 / 57.5	1,392 MB	15.7 ms
	Mean	**78.7 / 33.2**	**21,946 MB**	**1,430.6 ms**	**89.0 / 59.2**	**1,392 MB**	**15.7 ms**

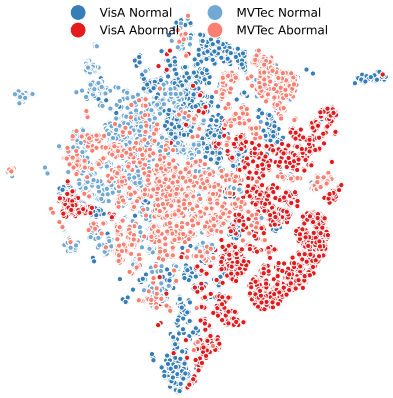
A4 Competing Methods

We compare state-of-the-art zero-shot AD methods, such as WinCLIP [22], AdaCLIP [4], AnomalyCLIP [55], Bayes-PFL [35], and AdaptCLIP [15], and few-shot AD methods, such as WinCLIP+ [22], InCtrl [57], AnomalyCLIP+, AdaptCLIP [15], FOUNDAD [52] and UniVAD [17], with our **UniADet**. It is worth noting that state-of-the-art UniVAD [17] uses a multi-model ensemble (*e.g.*, RAM, Grounding DINO, SAM, CLIP, and DINOv2), which results in slow inference speed and high GPU consumption. We compared it only on MVTec, VisA, and Real-IAD datasets. Following AdaptCLIP [15], we also report the few-shot performance of InCtrl [57] in both image-level and pixel-level, and extend zero-shot AnomalyCLIP [55] to few-shot AnomalyCLIP+. For more details, interested readers are refered to AdaptCLIP [15].

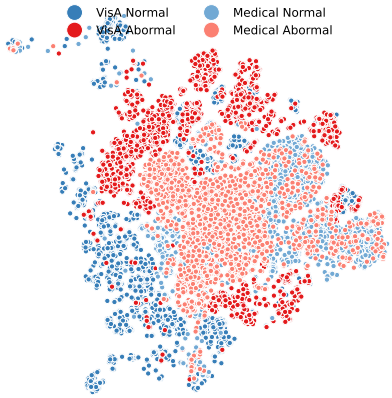
A5 Complete Results

In the main paper, we report category-averaged AUROC and AUPR for each dataset for image-level anomaly classification and pixel-level anomaly segmentation. Here, we provide more detailed and comprehensive results, including image-level anomaly classification in AUROC, AUPR, and $F1_{max}$, and pixel-level anomaly segmentation in AUROC, AUPR, $F1_{max}$, and AUPRO for 6 industrial benchmarks, MVTec (Tabs. A6, A7, A8), VisA (Tabs. A9, A10, A11), BTAD (Tabs. A12, A13, A14), DTD (Tabs. A15, A16, A17), KSDD (Tabs. A18, A19, A20), and Real-IAD (Tabs. A23, A24, A25, A26, A27, A28), and 8 medical benchmarks covering image anomaly classification (HeadCT, BrainMRI, Br35H in Tab. A21) and pixel anomaly segmentation (ISIC, ColonDB, ClinicDB, Endo, and Kvasir in Tab. A22). In addition, due to space limitations, we mainly compare the state-of-the-art few-shot methods with our **UniADet** with DINOv2R backbone in the main paper. Here, following previous works, we provide additional comparisons under 1-shot, 2-shot, and 4-shot in Tabs. A3, A4, and A5.

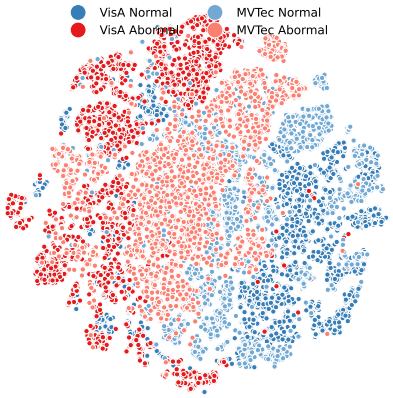
Furthermore, the main paper presents selected t-SNE visualizations; here we provide additional visualizations for all three backbones, CLIP ViT-L/14@336px, DINOv2R ViT-L/14 and DINOv3 ViT-L/16, as shown in Fig. A2. We can see that there are different manifolds between global image tokens and local patch tokens extracted from the same block, as well as between local patch tokens extracted from different blocks, regardless of whether the feature extractor is a visual-language CLIP model or self-supervised vision-only models, DINOv2R and DINOv3.



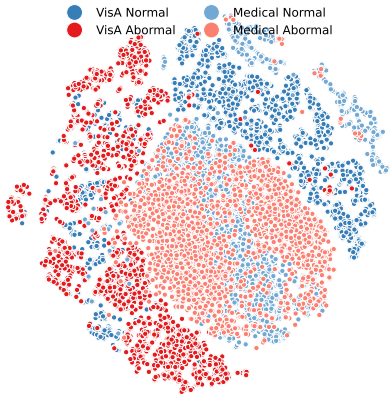
(a) VisA & MVTec on ViT-L-14@336px



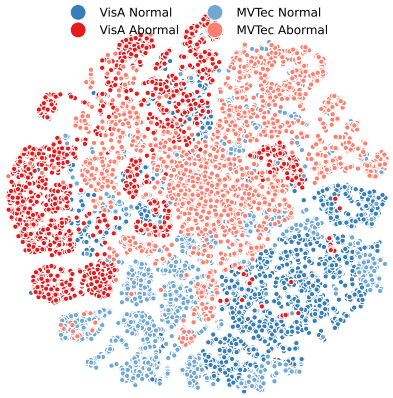
(b) VisA & Medical on ViT-L-14@336px



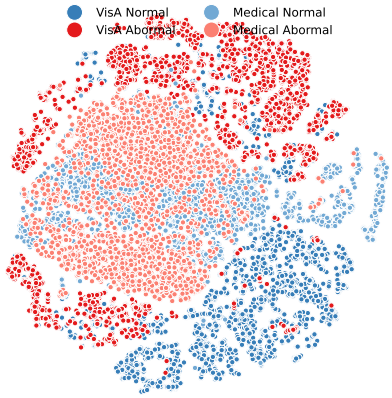
(c) VisA & MVTec on DINOv2R ViT-L/14



(d) VisA & Medical on DINOv2R ViT-L/14

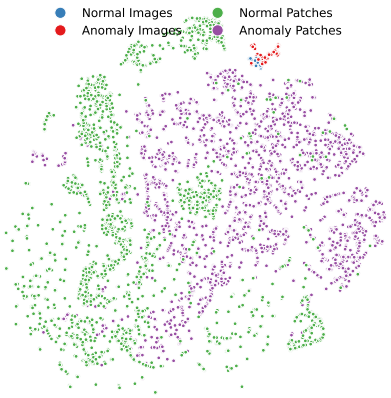


(e) VisA & MVTec on DINOv3 ViT-L/16

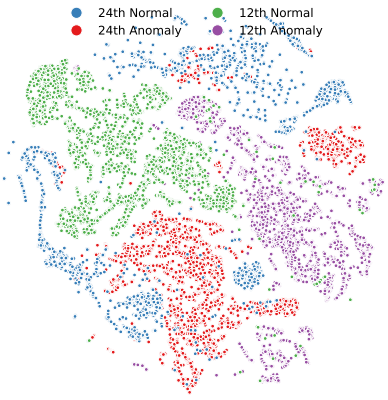


(f) VisA & Medical on DINOv3 ViT-L/16

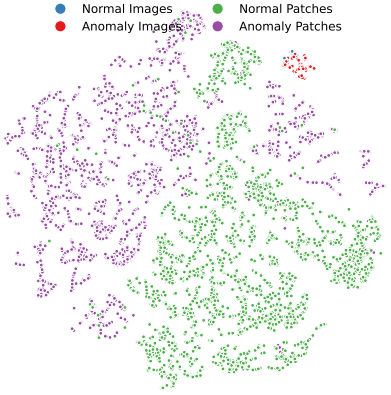
Figure A1: t-SNE visualizations of normal and anomalous features extracted using different foundation models, *e.g.*, ViT-L-14@336px, DINOv2R ViT-L/14, and DINOv3 ViT-L/16, on (a) (c) (e) on VisA and MVTec, and (b) (d) (f) on VisA and Medical.



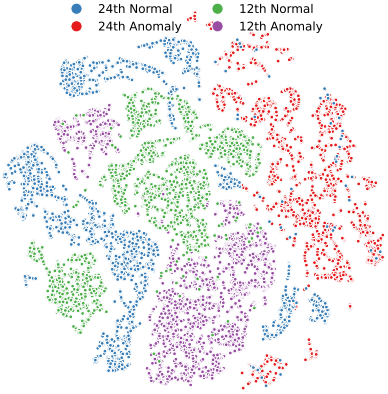
(a) Global vs. Local on ViT-L-14@336px



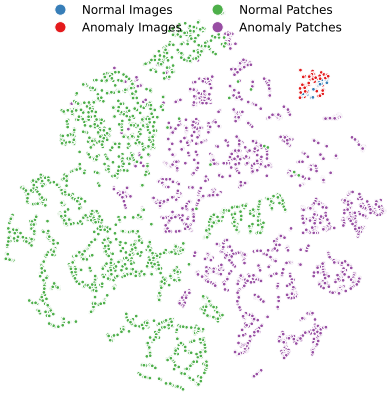
(b) Shallow vs. Deep on ViT-L-14@336px



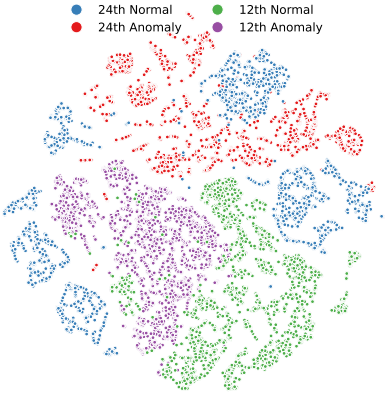
(c) Global vs. Local on DINOv2R ViT-L/14



(d) Shallow vs. Deep on DINOv2R ViT-L/14



(e) Global vs. Local on DINOv3 ViT-L/16



(f) Shallow vs. Deep on DINOv3 ViT-L/16

Figure A2: t-SNE visualization of features extracted using diverse foundation models, *e.g.*, ViT-L-14@336px, DINOv2R ViT-L/14, and DINOv3 ViT-L/16 on MVTec Test Set (Hazelnuts).

Table A3: Comparisons with state-of-the-art 1-shot AD methods in industrial domain. ✓ indicates language-free and ✗ means not language-free, the best mean results are marked in red, while the second-best are indicated in blue, and †, ‡ and ⋆ indicate CLIP ViT-L/14@336px, DINOv2-R ViT-L/16 and DINOv3 ViT-L/16, respectively, the same below.

Metric	Dataset	WinCLIP+[22] ✗	InCtrl[57] ✗	AnomalyCLIP+[55] ✗	AdaptCLIP[15] ✗	**UniADet†** ✓	**UniADet‡** ✓	**UniADet⋆** ✓
Image-Level (AUROC, AUPR)	MVTec	93.6±0.4 , 96.8±0.2	91.3±0.4 , 95.2±0.3	95.2±0.2 , 97.2±0.1	94.5±0.5 , 97.5±0.1	96.3±0.7 , 98.1±0.3	97.6±0.2 , 98.6±0.3	97.2±0.1 , 98.1±0.3
	VisA	80.0±2.4 , 81.7±1.5	83.2±2.4 , 84.1±1.5	86.1±0.7 , 87.7±1.1	90.5±1.2 , 92.3±0.9	89.3±1.9 , 90.6±1.5	95.2±0.8 , 95.6±0.6	95.8±0.8 , 95.5±0.5
	BTAD	84.4±1.5 , 80.5±3.3	88.5±0.4 , 83.4±9.3	88.5±0.8 , 74.2±1.5	93.4±0.0 , 95.8±0.9	96.6±0.3 , 97.4±0.4	96.4±0.3 , 97.7±0.4	97.1±0.5 , 98.5±0.2
	DTD	97.9±0.2 , 99.0±0.1	97.9±0.3 , 98.9±0.2	98.0±0.2 , 99.2±0.2	98.0±0.0 , 99.1±0.0	98.0±0.2 , 99.1±0.1	98.8±0.1 , 99.5±0.0	99.4±0.1 , 99.6±0.0
	KSDD	93.8±0.4 , 84.6±0.9	92.0±0.9 , 81.5±2.2	97.5±0.3 , 95.2±0.1	96.9±0.3 , 91.8±0.2	97.7±0.7 , 92.4±3.2	94.7±0.4 , 68.7±2.0	96.9±0.3 , 84.7±2.0
	Real-IAD	74.7±0.2 , 71.2±0.3	76.6±0.0 , 69.9±0.0	78.2±0.0 , 76.7±0.0	81.8±0.3 , 80.4±0.2	82.5±0.6 , 80.9±0.6	88.7±0.5 , 86.8±0.3	86.9±0.6 , 85.3±0.5
	Mean	87.4 , 85.6	88.3 , 85.5	90.6 , 88.4	92.5 , 92.8	93.4 , **93.1**	**95.2** , 91.2	**95.6** , **93.6**
Pixel-Level (AUROC, AUPR)	MVTec	93.4±0.2 , 38.3±0.8	94.6±0.2 , 47.8±1.1	92.8±0.0 , 40.8±0.1	94.3±0.1 , 53.7±0.9	95.6±0.3 , 55.7±1.3	97.1±0.1 , 63.1±0.5	97.4±0.1 , 66.5±0.9
	VisA	94.7±0.1 , 15.8±0.2	89.0±0.2 , 17.7±0.6	96.4±0.1 , 24.8±0.9	96.8±0.0 , 38.9±0.3	97.5±0.0 , 32.8±0.4	98.2±0.1 , 42.1±0.2	98.2±0.1 , 44.0±0.1
	BTAD	95.6±0.2 , 41.3±2.6	96.6±0.1 , 44.1±1.4	95.3±0.3 , 41.3±1.1	96.6±0.2 , 60.6±1.0	96.2±0.1 , 54.4±2.5	98.4±0.0 , 72.7±0.7	98.3±0.1 , 71.7±1.0
	DTD	96.5±0.1 , 47.8±0.9	98.6±0.1 , 64.3±0.5	97.6±0.1 , 67.4±0.4	97.4±0.0 , 76.9±0.1	97.4±0.1 , 73.4±0.9	98.9±0.1 , 82.4±0.3	99.1±0.0 , 85.8±0.2
	KSDD	97.6±0.1 , 19.2±0.3	97.8±0.2 , 26.7±0.7	98.6±0.1 , 47.5±0.5	98.2±0.1 , 57.8±1.2	98.0±0.0 , 45.0±5.8	99.4±0.3 , 34.9±2.7	99.3±0.1 , 45.2±3.7
	Real-IAD	95.0±0.0 , 13.9±0.2	95.4±0.0 , 19.1±0.0	96.5±0.0 , 27.9±0.0	97.1±0.0 , 36.6±0.1	98.0±0.0 , 35.3±0.2	98.9±0.0 , 48.4±0.4	98.6±0.1 , 47.6±0.5
	Mean	95.5 , 29.4	95.3 , 36.6	96.2 , 41.6	96.7 , **54.1**	97.1 , 49.4	**98.5** , **57.3**	**98.5** , **60.1**

Table A4: Comparisons with state-of-the-art 2-shot AD methods in industrial domain.

Metric	Dataset	WinCLIP+[22] ✗	InCtrl[57] ✗	AnomalyCLIP+[55] ✗	AdaptCLIP[15] ✗	**UniADet†** ✓	**UniADet‡** ✓	**UniADet⋆** ✓
Image-Level (AUROC, AUPR)	MVTec	94.5±1.0 , 97.3±0.5	91.8±0.9 , 95.5±0.7	95.4±0.1 , 97.3±0.1	95.7±0.6 , 97.9±0.2	97.0±0.6 , 98.3±0.2	98.0±0.3 , 98.9±0.1	97.6±0.3 , 98.6±0.2
	VisA	82.7±1.0 , 84.0±0.7	86.3±1.4 , 86.8±1.7	87.8±0.5 , 89.1±0.7	92.2±0.8 , 93.6±0.6	92.3±0.5 , 93.0±0.4	96.1±0.8 , 96.3±0.8	96.8±0.4 , 96.3±0.4
	BTAD	85.8±1.8 , 82.5±3.2	86.2±2.0 , 81.6±8.0	89.2±1.1 , 75.4±1.5	93.4±0.2 , 95.9±0.1	96.5±0.2 , 97.3±0.4	96.8±0.4 , 98.2±0.4	97.3±0.6 , 98.6±0.2
	DTD	98.1±0.2 , 99.1±0.1	98.3±0.2 , 99.1±0.3	98.2±0.1 , 99.3±0.1	97.4±0.0 , 99.2±0.0	98.0±0.1 , 99.1±0.1	99.2±0.1 , 99.6±0.0	99.6±0.1 , 99.7±0.1
	KSDD	93.8±0.2 , 84.5±0.6	91.6±0.9 , 81.0±2.5	97.9±0.2 , 95.6±0.2	97.2±0.0 , 92.4±0.4	97.7±0.6 , 93.3±1.3	95.3±0.2 , 71.2±1.2	97.2±0.4 , 87.5±0.9
	Real-IAD	76.1±0.1 , 72.7±0.1	78.5±0.0 , 71.7±0.0	78.3±0.0 , 76.9±0.0	82.9±0.2 , 81.5±0.1	83.6±0.2 , 82.0±0.3	89.0±0.7 , 86.7±1.4	87.2±1.2 , 85.0±2.2
	Mean	88.5 , 86.7	88.8 , 86.0	91.1 , 88.9	93.1 , 93.4	94.2 , **93.8**	**95.7** , 91.8	**96.0** , **94.3**
Pixel-Level (AUROC, AUPR)	MVTec	93.8±0.1 , 39.5±0.6	95.2±0.2 , 49.2±0.7	92.9±0.1 , 41.5±0.1	94.5±0.0 , 55.1±0.5	95.9±0.1 , 57.0±0.9	97.4±0.1 , 64.1±0.6	97.7±0.1 , 67.5±1.0
	VisA	95.1±0.1 , 17.2±0.8	89.8±0.2 , 18.5±0.2	96.7±0.1 , 26.2±0.7	97.1±0.0 , 40.7±0.6	97.8±0.0 , 35.8±1.0	98.4±0.0 , 44.2±0.3	98.3±0.0 , 45.3±0.4
	BTAD	95.7±0.1 , 42.8±1.3	96.7±0.1 , 44.2±0.8	95.5±0.2 , 41.9±0.6	96.7±0.1 , 61.0±0.6	96.4±0.1 , 54.5±2.1	98.4±0.0 , 73.1±0.5	98.4±0.0 , 71.8±1.3
	DTD	96.6±0.1 , 48.2±0.9	98.7±0.1 , 64.4±0.4	97.7±0.1 , 68.1±0.2	97.6±0.0 , 77.4±0.2	97.6±0.0 , 74.3±0.7	99.0±0.1 , 83.0±0.1	99.2±0.1 , 86.1±0.2
	KSDD	97.6±0.1 , 19.0±0.5	97.6±0.6 , 26.4±2.5	98.6±0.1 , 47.6±0.4	98.1±0.1 , 57.5±1.1	98.1±0.2 , 46.1±3.9	99.5±0.1 , 36.9±2.7	99.3±0.2 , 45.0±2.2
	Real-IAD	95.3±0.0 , 14.8±0.1	96.0±0.0 , 20.1±0.0	96.6±0.0 , 28.1±0.0	97.3±0.0 , 37.8±0.1	98.2±0.0 , 36.4±0.3	99.1±0.2 , 46.7±2.3	98.9±0.2 , 47.0±2.9
	Mean	95.7 , 30.3	95.7 , 37.1	96.3 , 42.2	96.9 , 54.9	**97.3** , 50.7	**98.6** , **58.0**	**98.6** , **60.5**

Table A5: Comparisons with state-of-the-art 4-shot AD methods in industrial domain.

Metric	Dataset	WinCLIP+[22] ✗	InCtrl[57] ✗	AnomalyCLIP+[55] ✗	AdaptCLIP[15] ✗	**UniADet†** ✓	**UniADet‡** ✓	**UniADet⋆** ✓
Image-Level (AUROC, AUPR)	MVTec	95.3±0.1 , 97.7±0.0	93.1±0.7 , 96.3±0.5	96.1±0.1 , 97.8±0.0	96.6±0.3 , 98.4±0.2	97.7±0.3 , 98.6±0.2	98.7±0.1 , 99.2±0.1	98.2±0.1 , 98.8±0.1
	VisA	84.3±0.6 , 85.5±0.9	87.8±0.2 , 88.0±0.3	88.8±0.5 , 90.1±0.7	93.1±0.2 , 94.3±0.2	93.3±0.2 , 93.8±0.1	96.9±0.3 , 96.8±0.3	97.1±0.1 , 96.6±0.1
	BTAD	87.8±0.8 , 88.1±1.4	67.5±2.4 , 80.9±1.7	90.5±1.2 , 77.5±3.1	93.3±0.3 , 96.4±0.1	96.7±0.2 , 97.8±0.4	96.9±0.3 , 98.4±0.3	97.5±0.3 , 98.8±0.2
	DTD	98.2±0.0 , 99.2±0.1	97.7±0.1 , 98.3±0.4	98.4±0.1 , 99.4±0.1	98.5±0.1 , 99.3±0.0	98.4±0.4 , 99.2±0.2	99.7±0.1 , 99.7±0.0	99.7±0.1 , 99.7±0.1
	KSDD	94.0±0.2 , 84.9±0.5	91.6±0.9 , 84.6±1.7	97.8±0.1 , 95.0±0.2	97.0±0.2 , 91.7±0.9	97.3±0.8 , 91.3±1.8	95.3±0.6 , 71.7±1.9	96.7±0.2 , 86.5±1.7
	Real-IAD	77.0±0.0 , 73.6±0.1	81.8±0.0 , 75.6±0.0	78.4±0.0 , 77.1±0.0	83.9±0.2 , 82.6±0.0	84.3±0.2 , 82.6±0.2	90.3±0.2 , 88.4±0.3	88.5±0.2 , 87.0±0.2
	Mean	89.4 , 88.2	86.6 , 87.3	91.7 , 89.5	93.7 , 93.8	**94.6** , **93.9**	**96.3** , 92.4	**96.3** , **94.6**
Pixel-Level (AUROC, AUPR)	MVTec	94.2±0.2 , 41.2±0.9	95.8±0.2 , 50.9±0.3	93.2±0.1 , 42.4±0.0	94.8±0.1 , 57.2±0.8	96.3±0.1 , 58.8±0.5	97.7±0.1 , 65.4±0.2	97.9±0.0 , 69.0±0.4
	VisA	95.1±0.2 , 18.1±1.3	90.2±0.2 , 19.2±0.6	96.9±0.1 , 27.5±1.1	97.3±0.0 , 41.8±0.6	98.0±0.1 , 36.7±1.3	98.6±0.1 , 45.2±0.6	98.4±0.1 , 45.5±0.6
	BTAD	95.9±0.1 , 44.0±0.4	96.8±0.0 , 44.0±0.2	95.7±0.1 , 45.8±3.0	96.8±0.0 , 62.3±0.3	96.6±0.1 , 56.5±1.4	98.4±0.1 , 73.9±0.5	98.4±0.0 , 72.7±1.0
	DTD	96.8±0.1 , 49.3±0.1	98.7±0.1 , 64.9±0.3	97.8±0.0 , 68.5±0.2	97.8±0.1 , 78.2±0.2	97.7±0.1 , 75.2±0.3	99.2±0.1 , 83.8±0.1	99.3±0.1 , 86.7±0.1
	KSDD	97.5±0.2 , 19.1±0.7	97.5±0.3 , 26.0±1.4	98.6±0.1 , 46.4±0.7	98.0±0.1 , 56.4±1.4	98.1±0.2 , 42.5±5.7	99.4±0.2 , 35.2±5.3	99.3±0.2 , 44.0±5.7
	Real-IAD	95.5±0.0 , 15.4±0.2	96.4±0.0 , 21.0±0.0	96.7±0.0 , 28.2±0.0	97.4±0.0 , 39.1±0.3	98.3±0.1 , 37.2±0.2	99.1±0.1 , 48.5±1.2	98.9±0.1 , 49.8±0.2
	Mean	95.8 , 31.2	95.9 , 37.7	96.5 , 43.1	97.0 , **55.8**	**97.5** , 51.2	**98.7** , **58.7**	**98.7** , **61.3**

Table A6: Image anomaly classification and pixel anomaly segmentation results on **MVTec** with zero-shot and few-shot **UniADet**†.

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
0-shot	carpet	99.9	99.8	99.4	99.2	67.6	63.6	98.3
	bottle	93.7	98.2	92.2	91.4	55.5	52.7	84.7
	hazelnut	95.4	97.2	93.8	96.7	52.6	51.5	93.0
	leather	100.0	99.9	100.0	99.4	53.5	51.5	99.1
	cable	76.7	82.6	79.5	80.2	21.2	27.7	69.3
	capsule	92.7	98.5	92.7	96.4	32.4	36.7	91.2
	grid	99.3	99.6	98.2	98.0	40.1	40.1	90.9
	pill	85.3	96.7	92.5	86.9	25.3	26.6	93.7
	transistor	84.6	83.4	79.5	71.3	12.2	18.5	56.5
	metal_nut	88.6	97.2	92.1	77.0	31.0	35.3	81.7
	screw	84.2	94.0	88.4	99.1	39.6	40.4	95.8
	toothbrush	90.6	95.9	93.5	94.1	26.9	31.9	92.4
	zipper	96.0	98.9	95.0	97.3	58.4	57.1	91.7
	tile	99.8	99.8	98.8	93.9	66.3	68.0	89.6
	wood	99.2	99.6	98.3	96.5	59.6	59.8	95.4
	Mean	92.4	96.1	92.9	91.8	42.8	44.1	88.2
1-shot	carpet	100.0±0.1	99.9±0.1	99.8±0.3	99.0±0.1	70.8±0.5	69.6±0.2	97.5±0.1
	bottle	99.3±0.2	99.6±0.1	97.7±0.0	97.0±0.1	78.9±0.5	73.7±0.4	94.0±0.3
	hazelnut	99.0±0.3	99.3±0.1	96.5±0.7	98.9±0.1	69.0±1.4	66.6±1.4	96.5±0.2
	leather	100.0±0.0	99.9±0.0	100.0±0.0	99.5±0.1	55.5±0.4	56.5±1.7	98.9±0.4
	cable	93.9±0.3	96.2±0.4	90.1±0.8	93.1±0.3	54.6±1.0	56.0±0.8	87.3±0.2
	capsule	88.7±7.0	97.4±1.7	94.7±1.5	97.3±0.5	31.1±13.0	35.4±10.2	94.0±1.9
	grid	98.7±0.2	99.4±0.1	97.4±0.8	97.9±0.2	42.5±1.5	46.7±0.9	93.1±0.5
	pill	96.4±1.1	99.2±0.3	96.7±0.2	95.3±0.6	56.8±1.9	58.2±1.6	96.9±0.1
	transistor	93.5±1.6	90.4±2.3	85.0±3.3	80.9±1.7	35.4±4.3	39.8±4.0	64.7±1.4
	metal_nut	99.5±0.2	99.8±0.1	98.6±0.6	91.0±1.0	59.5±3.0	57.9±2.9	91.2±0.7
	screw	81.4±2.9	92.6±1.3	88.5±1.2	97.3±0.2	25.5±3.5	32.0±4.0	89.5±1.0
	toothbrush	97.8±0.9	98.8±0.3	95.7±1.7	98.8±0.2	62.6±1.7	63.9±1.5	93.6±1.3
	zipper	97.2±1.1	99.2±0.3	95.6±0.9	96.4±0.3	56.1±1.2	54.9±1.1	89.6±0.6
	tile	99.9±0.1	99.8±0.0	99.6±0.3	95.9±0.1	71.0±0.3	69.9±0.2	93.2±0.1
	wood	99.6±0.0	99.7±0.0	98.9±0.5	95.9±0.4	66.9±1.1	65.6±0.5	95.5±0.1
	Mean	96.3±0.7	98.1±0.3	95.6±0.2	95.6±0.3	55.7±1.3	56.4±1.3	91.7±0.4
2-shot	carpet	100.0±0.1	99.9±0.0	99.4±0.0	99.1±0.0	70.7±0.5	69.5±0.2	97.6±0.1
	bottle	99.3±0.2	99.6±0.1	97.4±0.5	97.2±0.1	79.5±0.6	74.2±0.4	94.4±0.1
	hazelnut	99.8±0.2	99.6±0.1	98.8±0.8	98.9±0.0	69.4±1.2	67.0±0.2	96.8±0.1
	leather	100.0±0.0	99.9±0.0	100.0±0.0	99.5±0.0	55.4±0.5	56.5±1.1	98.9±0.4
	cable	93.4±1.1	95.9±0.7	89.9±0.5	93.4±0.2	55.4±0.5	56.5±0.7	88.4±0.4
	capsule	92.0±7.9	98.1±1.9	95.7±1.6	97.7±0.6	38.8±14.3	41.5±10.4	95.4±1.9
	grid	99.4±0.6	99.6±0.2	98.5±1.4	98.2±0.3	42.3±1.6	45.9±0.5	93.7±0.6
	pill	96.4±0.7	99.2±0.2	96.6±0.2	95.4±0.5	57.1±1.5	58.7±0.9	97.0±0.2
	transistor	94.6±0.4	92.0±0.9	86.7±1.9	81.4±1.5	35.7±4.0	40.1±3.7	65.2±1.3
	metal_nut	99.7±0.1	99.8±0.0	98.9±0.6	92.2±0.5	63.1±1.6	61.7±1.9	92.2±0.7
	screw	84.2±1.1	93.2±1.3	90.2±1.2	97.9±0.3	27.1±1.6	34.3±0.6	91.5±1.5
	toothbrush	98.0±1.2	98.8±0.5	96.8±1.6	99.0±0.2	63.9±2.9	64.8±2.7	94.5±1.7
	zipper	98.1±0.7	99.5±0.2	96.2±1.2	96.9±0.2	57.9±1.2	56.8±1.0	90.8±0.6
	tile	99.9±0.1	99.8±0.0	99.6±0.3	95.9±0.1	70.7±0.2	69.7±0.0	93.1±0.1
	wood	99.6±0.1	99.7±0.0	98.9±0.5	96.0±0.4	67.8±0.7	65.7±0.4	95.5±0.2
	Mean	97.0±0.6	98.3±0.2	96.3±0.1	95.9±0.1	57.0±0.9	57.5±0.8	92.3±0.3
4-shot	carpet	100.0±0.1	99.9±0.0	99.4±0.0	99.1±0.0	71.2±0.7	69.6±0.2	97.6±0.0
	bottle	99.7±0.3	99.7±0.1	98.7±0.5	97.4±0.1	79.9±0.5	74.7±0.4	94.6±0.1
	hazelnut	99.6±0.6	99.5±0.3	98.6±1.3	99.0±0.1	70.2±0.6	68.2±0.9	96.9±0.4
	leather	100.0±0.0	99.9±0.0	100.0±0.0	99.5±0.0	54.7±0.5	56.3±0.8	98.9±0.3
	cable	94.5±0.7	96.5±0.3	89.8±1.2	93.8±0.2	57.0±0.4	57.7±0.3	88.7±0.4
	capsule	97.7±1.7	99.4±0.4	97.6±1.3	98.3±0.2	48.0±1.9	48.0±1.5	96.9±0.3
	grid	99.5±0.4	99.7±0.1	98.8±0.5	98.2±0.1	41.8±0.8	45.7±0.8	93.9±0.4
	pill	96.3±0.3	99.2±0.1	96.5±0.0	95.6±0.1	57.8±0.6	59.3±0.4	97.1±0.1
	transistor	95.3±1.2	93.1±1.7	87.1±0.6	83.1±0.7	37.7±1.9	42.2±1.8	66.7±0.9
	metal_nut	99.7±0.3	99.9±0.1	99.3±0.3	93.3±0.6	66.5±2.3	65.1±2.7	93.1±0.8
	screw	87.7±3.5	93.9±1.7	91.8±0.8	98.6±0.4	35.0±7.0	43.0±6.9	93.9±1.1
	toothbrush	98.2±0.9	98.9±0.4	96.8±1.6	99.2±0.2	65.3±2.7	65.8±2.5	95.5±1.0
	zipper	98.7±0.9	99.6±0.2	97.1±1.4	97.2±0.1	58.8±0.8	57.6±0.5	91.6±0.2
	tile	100.0±0.0	99.8±0.0	99.8±0.3	96.0±0.1	70.8±0.6	69.7±0.2	93.1±0.1
	wood	99.5±0.2	99.7±0.0	98.3±0.0	96.1±0.2	68.2±0.7	65.7±0.3	95.4±0.2
	Mean	97.7±0.3	98.6±0.2	96.6±0.2	96.3±0.1	58.8±0.5	59.3±0.4	92.9±0.1

Table A7: Image anomaly classification and pixel anomaly segmentation results on **MVTec** with zero-shot and few-shot **UniADet**[‡].

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
0-shot	carpet	98.7	99.5	97.3	99.5	75.9	70.1	98.9
	bottle	99.7	99.7	98.4	93.9	68.5	62.9	90.0
	hazelnut	96.5	98.0	92.6	97.0	58.1	55.0	91.4
	leather	99.9	99.8	98.9	99.4	50.9	49.4	99.2
	cable	81.6	88.9	80.4	81.5	23.1	28.3	73.7
	capsule	94.8	98.9	95.5	96.3	34.2	34.9	91.6
	grid	100.0	99.8	100.0	99.2	48.7	46.3	97.0
	pill	85.2	97.3	91.6	87.5	32.2	33.6	95.4
	transistor	87.6	87.6	83.5	69.1	14.1	17.7	55.6
	metal_nut	75.6	94.7	89.4	77.3	32.6	36.5	83.1
	screw	89.4	96.6	88.4	98.9	67.4	64.0	95.7
	toothbrush	95.6	98.5	96.6	95.7	37.5	39.1	95.5
	zipper	99.6	99.8	98.8	98.8	77.3	71.4	96.5
	tile	98.8	99.4	96.5	95.5	75.6	71.9	91.2
	wood	99.6	99.7	99.2	97.2	66.8	64.2	97.0
	Mean	93.5	97.2	93.8	92.4	50.9	49.7	90.1
1-shot	carpet	99.4±0.3	99.7±0.1	98.3±0.6	99.5±0.0	75.4±0.6	69.7±0.4	98.8±0.1
	bottle	99.9±0.1	99.8±0.0	99.7±0.5	98.9±0.1	86.9±0.4	79.4±0.5	97.3±0.3
	hazelnut	100.0±0.0	99.7±0.0	99.5±0.4	99.4±0.0	78.3±0.8	76.6±0.8	95.5±0.9
	leather	100.0±0.0	99.9±0.0	100.0±0.0	99.2±0.2	49.0±4.4	47.5±3.7	99.0±0.2
	cable	94.0±0.6	97.0±0.1	91.3±0.7	96.5±0.1	65.1±1.5	67.6±1.5	92.3±0.3
	capsule	94.6±1.3	98.9±0.4	94.9±0.2	97.8±0.2	44.2±4.5	45.2±5.1	94.2±0.5
	grid	100.0±0.0	99.8±0.0	100.0±0.0	99.4±0.1	54.4±0.2	53.6±0.5	97.4±0.1
	pill	97.3±0.9	99.5±0.2	96.7±0.8	92.8±0.5	52.5±2.1	53.0±2.1	96.9±0.1
	transistor	92.8±2.4	90.5±3.8	86.2±4.6	90.5±1.5	47.3±2.4	48.3±2.4	73.8±1.7
	metal_nut	95.2±0.6	98.8±0.1	94.9±0.7	95.6±0.4	71.7±2.6	72.3±2.8	94.1±0.2
	screw	90.9±1.6	97.1±0.5	89.7±0.5	98.1±0.4	58.0±2.9	58.2±1.9	93.0±0.9
	toothbrush	99.4±0.6	99.3±0.2	97.9±1.8	99.2±0.2	60.6±1.5	62.1±2.4	97.7±0.1
	zipper	99.9±0.1	99.9±0.0	99.7±0.2	96.2±0.3	59.9±2.6	55.3±2.7	91.8±0.5
	tile	99.7±0.0	99.7±0.0	99.4±0.0	97.9±0.0	76.9±0.1	72.4±0.1	95.6±0.1
	wood	99.9±0.1	99.8±0.0	99.5±0.5	95.4±0.2	65.9±0.5	64.4±0.1	96.1±0.2
	Mean	97.6±0.2	98.6±0.3	96.5±0.5	97.1±0.1	63.1±0.5	61.7±0.4	94.2±0.2
2-shot	carpet	99.6±0.1	99.7±0.1	98.9±0.0	99.5±0.0	74.8±0.2	69.3±0.2	98.7±0.0
	bottle	99.9±0.1	99.8±0.1	99.7±0.5	99.0±0.1	87.7±0.3	80.0±0.5	97.5±0.3
	hazelnut	100.0±0.0	99.7±0.0	99.8±0.4	99.4±0.1	78.5±0.9	76.6±0.7	96.0±0.2
	leather	100.0±0.0	99.9±0.0	100.0±0.0	99.2±0.1	49.1±3.5	47.5±2.6	99.1±0.2
	cable	94.8±1.3	97.4±0.5	92.5±0.5	97.0±0.3	67.0±0.9	68.6±0.7	93.5±0.8
	capsule	93.9±1.7	98.7±0.4	94.7±0.8	98.0±0.2	47.2±5.8	48.4±6.2	94.5±0.7
	grid	100.0±0.0	99.8±0.0	100.0±0.0	99.5±0.0	55.4±0.4	54.7±0.4	97.6±0.1
	pill	97.3±0.7	99.5±0.1	97.2±0.6	93.3±0.7	53.3±2.1	53.6±1.3	97.0±0.1
	transistor	94.9±0.4	93.8±0.6	89.7±0.5	91.4±1.1	49.0±2.1	49.3±1.6	75.5±1.4
	metal_nut	96.9±1.4	99.2±0.3	96.5±1.4	96.2±0.3	73.8±1.1	75.5±0.8	94.7±0.5
	screw	92.6±2.0	97.7±0.6	91.0±1.3	98.5±0.4	59.3±1.3	58.9±1.4	94.5±0.7
	toothbrush	99.9±0.2	99.5±0.0	99.5±0.9	99.2±0.1	62.7±2.0	62.8±2.0	98.0±0.0
	zipper	100.0±0.0	99.9±0.0	99.9±0.2	96.6±0.4	60.8±1.2	56.1±1.0	92.4±0.6
	tile	99.7±0.1	99.7±0.1	99.4±0.0	97.9±0.1	76.2±0.8	71.9±0.4	95.4±0.2
	wood	99.9±0.1	99.8±0.0	99.5±0.5	96.1±0.2	67.4±0.3	65.2±0.2	96.4±0.2
	Mean	98.0±0.3	98.9±0.1	97.2±0.2	97.4±0.1	64.1±0.6	62.6±0.5	94.7±0.1
4-shot	carpet	99.7±0.0	99.8±0.0	99.2±0.3	99.5±0.0	75.2±0.5	69.5±0.3	98.7±0.0
	bottle	100.0±0.0	99.8±0.0	100.0±0.0	99.1±0.0	88.4±0.2	80.7±0.2	97.7±0.2
	hazelnut	100.0±0.0	99.7±0.0	99.8±0.4	99.5±0.0	80.0±1.1	77.6±0.5	96.5±0.6
	leather	100.0±0.0	99.9±0.0	100.0±0.0	99.2±0.1	48.5±3.1	46.8±2.4	99.0±0.2
	cable	95.1±0.3	97.6±0.1	93.5±0.7	97.0±0.2	67.5±0.3	68.7±0.3	93.5±0.4
	capsule	98.2±0.4	99.6±0.1	98.2±0.5	98.4±0.2	53.6±1.0	54.0±0.8	95.4±0.4
	grid	100.0±0.0	99.8±0.0	100.0±0.0	99.5±0.0	56.1±0.5	55.6±0.2	97.8±0.2
	pill	97.8±0.1	99.6±0.1	97.9±0.0	93.3±0.3	53.3±1.3	53.9±0.9	97.1±0.1
	transistor	96.6±1.2	95.6±1.4	91.0±2.1	93.1±0.5	53.0±1.0	52.1±0.7	78.6±1.2
	metal_nut	99.9±0.1	99.9±0.0	99.5±0.6	96.7±0.5	75.7±1.6	78.4±2.6	95.3±0.5
	screw	93.9±0.4	98.0±0.2	91.7±1.1	98.9±0.1	61.2±2.8	59.0±1.0	95.7±0.2
	toothbrush	100.0±0.0	99.5±0.0	100.0±0.0	99.3±0.1	64.6±2.2	63.7±1.9	98.1±0.2
	zipper	100.0±0.0	99.9±0.0	99.9±0.2	96.8±0.1	61.0±1.0	56.4±1.0	92.8±0.2
	tile	99.9±0.2	99.8±0.1	99.8±0.3	97.9±0.0	76.0±0.5	72.0±0.3	95.3±0.1
	wood	100.0±0.1	99.8±0.0	99.7±0.5	96.4±0.3	67.4±0.9	64.9±0.4	96.5±0.2
	Mean	98.7±0.1	99.2±0.1	98.0±0.2	97.7±0.1	65.4±0.2	63.6±0.2	95.2±0.1

Table A8: Image anomaly classification and pixel anomaly segmentation results on **MVTec** with zero-shot and few-shot **UniADet***.

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
0-shot	carpet	99.8	99.8	98.9	99.4	76.1	70.4	98.9
	bottle	98.3	99.5	97.6	95.8	73.2	66.0	92.7
	hazelnut	97.7	98.6	95.6	97.4	62.0	58.5	91.7
	leather	100.0	99.9	100.0	99.5	58.0	52.9	99.4
	cable	81.4	88.0	82.8	81.0	17.8	24.7	72.7
	capsule	96.6	99.4	96.7	96.5	36.3	36.4	92.2
	grid	99.9	99.7	99.1	99.3	53.4	48.8	96.8
	pill	83.6	96.9	91.6	86.9	30.4	32.1	93.3
	transistor	83.8	83.8	76.1	72.2	13.3	19.1	61.5
	metal_nut	80.2	95.6	89.4	83.8	39.5	42.0	86.5
	screw	95.2	98.4	92.6	99.2	66.5	63.1	95.7
	toothbrush	94.4	98.2	94.7	95.2	35.8	37.5	92.7
	zipper	99.6	99.8	98.3	99.2	80.4	75.7	97.6
	tile	99.5	99.7	97.6	95.4	74.7	71.0	91.4
	wood	99.8	99.8	99.2	97.4	72.8	67.5	97.5
	Mean	94.0	97.1	94.0	93.2	52.7	51.1	90.7
1-shot	carpet	100.0±0.1	99.9±0.1	99.6±0.3	99.6±0.1	79.2±0.1	74.9±0.1	99.1±0.0
	bottle	99.5±0.1	99.7±0.0	99.2±0.0	98.4±0.1	85.8±0.3	78.4±0.1	97.0±0.1
	hazelnut	100.0±0.0	99.7±0.0	100.0±0.0	99.4±0.1	85.5±0.9	80.2±0.9	93.7±0.0
	leather	100.0±0.0	99.9±0.0	100.0±0.0	99.6±0.1	64.3±3.3	58.1±3.0	99.4±0.1
	cable	95.6±0.1	97.9±0.1	93.8±1.2	96.7±0.1	67.4±2.0	67.6±2.2	92.2±0.4
	capsule	89.9±4.0	97.8±0.9	94.4±1.1	97.9±0.2	40.4±9.9	41.8±7.6	94.4±0.6
	grid	100.0±0.0	99.8±0.0	100.0±0.0	99.6±0.0	62.9±0.6	59.7±0.8	98.0±0.3
	pill	96.9±0.1	99.4±0.1	96.9±0.5	93.1±0.6	47.8±1.3	46.3±1.6	96.6±0.3
	transistor	87.0±3.5	82.3±6.1	78.1±4.0	90.9±1.4	44.2±1.2	44.1±1.0	78.1±2.2
	metal_nut	100.0±0.0	99.9±0.0	100.0±0.0	95.1±0.3	72.0±1.9	70.1±1.9	93.5±0.1
	screw	89.2±1.5	96.6±0.5	88.2±0.8	97.2±0.3	48.7±2.6	52.4±2.2	88.9±0.9
	toothbrush	99.9±0.2	99.5±0.0	99.5±0.9	99.3±0.2	69.0±2.1	67.2±2.7	96.6±0.4
	zipper	99.6±0.6	99.8±0.2	99.2±1.1	98.1±0.3	71.8±0.1	65.9±0.5	95.9±0.3
	tile	99.9±0.1	99.8±0.0	99.6±0.3	98.6±0.1	81.8±0.1	77.4±0.2	96.8±0.1
	wood	99.7±0.1	99.7±0.0	98.9±0.5	97.2±0.3	77.0±1.1	70.4±0.4	97.4±0.2
	Mean	97.2±0.1	98.1±0.3	96.5±0.1	97.4±0.1	66.5±0.9	63.6±0.7	94.5±0.2
2-shot	carpet	100.0±0.1	99.9±0.1	99.4±0.0	99.7±0.1	79.1±0.2	74.8±0.1	99.1±0.0
	bottle	99.4±0.5	99.7±0.1	99.2±0.0	98.6±0.1	86.7±0.1	79.0±0.1	97.3±0.0
	hazelnut	100.0±0.0	99.7±0.0	100.0±0.0	99.4±0.0	84.7±0.9	79.5±1.1	94.0±0.6
	leather	100.0±0.0	99.9±0.0	100.0±0.0	99.6±0.1	64.5±2.5	58.2±2.4	99.4±0.1
	cable	95.3±0.3	97.6±0.1	93.0±0.4	96.9±0.2	67.8±0.4	67.6±0.8	92.8±0.6
	capsule	92.8±4.6	98.4±1.0	95.3±1.3	98.2±0.3	46.8±10.8	46.6±8.0	95.2±0.9
	grid	100.0±0.0	99.8±0.0	100.0±0.0	99.6±0.0	62.7±0.8	59.5±0.8	97.9±0.1
	pill	96.8±0.4	99.4±0.1	96.9±0.7	93.4±0.6	48.0±1.3	46.7±1.1	96.6±0.1
	transistor	90.8±2.3	88.4±2.0	81.2±2.4	91.7±1.3	46.1±1.6	46.0±1.3	79.8±2.8
	metal_nut	100.0±0.0	99.9±0.0	100.0±0.0	95.9±0.3	74.8±0.6	73.4±1.1	94.2±0.6
	screw	90.3±1.4	96.8±0.4	89.8±1.6	97.9±0.7	50.7±3.4	53.3±3.4	91.3±1.7
	toothbrush	99.8±0.3	99.5±0.1	99.5±0.9	99.4±0.1	69.9±2.4	67.5±1.9	97.0±0.5
	zipper	99.9±0.2	99.9±0.1	99.7±0.5	98.3±0.2	72.1±0.2	66.0±0.2	96.2±0.3
	tile	100.0±0.1	99.8±0.0	99.6±0.3	98.5±0.0	81.4±0.3	77.2±0.0	96.7±0.1
	wood	99.7±0.1	99.7±0.1	98.6±0.5	97.3±0.2	77.6±0.7	70.7±0.2	97.5±0.2
	Mean	97.6±0.3	98.6±0.2	96.8±0.1	97.7±0.1	67.5±1.0	64.4±0.8	95.0±0.0
4-shot	carpet	99.9±0.1	99.9±0.1	99.4±0.0	99.7±0.0	79.3±0.3	74.9±0.3	99.1±0.0
	bottle	99.3±0.4	99.7±0.1	99.2±0.0	98.8±0.1	87.2±0.2	79.3±0.1	97.5±0.1
	hazelnut	100.0±0.0	99.7±0.0	99.8±0.4	99.5±0.1	85.3±0.8	79.9±0.7	94.7±0.7
	leather	100.0±0.0	99.9±0.0	100.0±0.0	99.6±0.0	63.1±3.1	56.9±2.6	99.3±0.1
	cable	95.2±0.8	97.6±0.4	93.6±0.5	96.9±0.1	68.0±0.1	67.6±0.5	92.6±0.3
	capsule	98.2±0.5	99.6±0.1	97.7±0.4	98.6±0.1	55.9±1.2	53.3±1.0	96.3±0.4
	grid	100.0±0.0	99.8±0.0	100.0±0.0	99.6±0.0	62.7±0.6	59.7±0.6	98.1±0.3
	pill	96.7±0.2	99.4±0.1	96.6±0.8	93.6±0.2	47.9±0.6	46.8±0.6	96.7±0.1
	transistor	90.8±1.0	89.8±0.6	84.0±1.1	93.4±0.5	48.8±0.5	48.1±0.4	83.9±1.8
	metal_nut	100.0±0.0	99.9±0.0	100.0±0.0	96.5±0.6	77.0±2.3	76.4±3.1	94.9±0.8
	screw	93.2±1.1	97.7±0.4	92.6±0.6	98.5±0.1	57.1±4.9	57.4±3.1	93.1±0.8
	toothbrush	100.0±0.0	99.5±0.0	100.0±0.0	99.5±0.1	71.3±2.5	68.4±2.1	97.3±0.5
	zipper	100.0±0.0	99.9±0.0	100.0±0.0	98.4±0.1	72.3±0.4	66.0±0.3	96.4±0.2
	tile	100.0±0.0	99.8±0.0	100.0±0.0	98.5±0.1	81.5±0.2	77.2±0.1	96.6±0.1
	wood	99.7±0.1	99.7±0.0	98.9±0.5	97.4±0.2	77.5±0.8	70.7±0.3	97.5±0.2
	Mean	98.2±0.1	98.8±0.1	97.5±0.1	97.9±0.0	69.0±0.4	65.5±0.4	95.6±0.2

Table A9: Image anomaly classification and pixel anomaly segmentation results on **VisA** with zero-shot and few-shot **UniADet**†.

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	**I-AUPR**	**I-F1$_{max}$**	**P-AUROC**	**P-AUPR**	**P-F1$_{max}$**	**P-AUPRO**
0-shot	candle	94.5	95.0	86.9	99.3	39.5	46.9	95.6
	capsules	94.3	96.6	91.5	99.0	51.4	53.1	94.8
	cashew	89.4	95.4	86.0	90.3	16.3	21.8	92.8
	chewinggum	97.3	98.9	96.4	99.6	83.2	76.6	95.3
	fryum	97.0	98.5	93.1	94.7	24.2	31.1	93.0
	macaroni1	89.9	90.6	82.0	99.5	30.2	38.5	95.8
	macaroni2	73.9	72.5	73.3	98.8	8.7	17.3	89.3
	pcb1	85.6	86.6	79.6	93.0	7.7	12.3	89.7
	pcb2	74.0	72.7	72.8	91.9	11.4	19.2	78.3
	pcb3	71.6	74.8	69.7	91.2	9.5	17.9	81.6
	pcb4	88.5	88.4	84.0	96.1	33.8	38.0	87.8
	pipe_fryum	99.4	99.6	98.5	95.7	19.9	26.3	98.0
	Mean	88.0	89.1	84.5	95.8	28.0	33.3	91.0
1-shot	candle	92.9±0.2	92.2±0.3	87.5±1.1	99.3±0.1	29.9±3.3	42.3±1.2	96.8±0.3
	capsules	94.2±1.0	96.7±0.6	90.0±0.5	97.3±0.2	37.0±1.5	43.3±1.6	93.4±0.1
	cashew	92.5±1.3	96.5±0.7	90.0±0.5	97.7±0.2	53.3±3.1	58.8±2.2	94.2±2.2
	chewinggum	97.7±0.3	99.0±0.1	96.0±0.4	99.5±0.0	68.5±1.3	71.2±0.7	95.4±0.3
	fryum	98.4±0.2	99.2±0.1	96.1±0.5	96.3±0.5	35.2±2.3	42.9±2.6	93.2±0.3
	macaroni1	93.6±0.6	94.0±0.3	87.5±1.2	99.4±0.1	21.5±2.6	30.6±1.9	93.6±0.3
	macaroni2	76.4±1.7	78.9±0.4	72.4±2.1	97.1±0.3	9.4±0.9	18.3±1.6	79.3±1.9
	pcb1	80.4±19.0	81.4±13.1	80.1±9.8	98.3±0.2	34.6±3.6	41.5±1.7	92.0±2.7
	pcb2	77.9±8.2	76.8±9.2	74.9±3.7	95.3±0.3	14.8±4.7	25.2±6.4	81.3±0.9
	pcb3	82.2±4.3	84.8±3.0	75.8±4.0	95.1±0.1	30.0±3.6	39.1±3.4	84.7±0.3
	pcb4	86.0±16.4	87.4±10.6	83.3±14.3	96.7±0.2	24.3±0.8	32.7±1.0	88.4±1.2
	pipe_fryum	99.8±0.3	99.7±0.1	99.3±0.8	98.2±0.2	35.3±2.1	46.4±2.0	97.6±0.1
	Mean	89.3±1.9	90.6±1.5	86.1±1.6	97.5±0.0	32.8±0.4	41.0±0.8	90.8±0.5
2-shot	candle	93.7±0.5	93.1±0.6	88.2±0.5	99.3±0.1	29.8±2.0	42.2±0.5	97.0±0.1
	capsules	94.9±1.6	97.0±1.1	90.5±2.3	97.8±0.2	40.9±1.2	46.5±1.1	94.0±1.1
	cashew	93.1±0.4	96.7±0.2	89.9±1.0	98.0±0.1	55.2±0.7	60.0±0.6	94.0±1.4
	chewinggum	97.8±0.1	99.1±0.1	96.3±0.2	99.5±0.1	69.2±1.0	71.3±0.4	95.4±0.1
	fryum	98.6±0.1	99.3±0.1	96.3±0.3	96.6±0.3	36.7±0.7	44.8±0.9	93.6±0.5
	macaroni1	93.1±1.4	93.7±0.8	86.0±3.3	99.5±0.1	22.1±3.4	30.9±2.1	94.0±0.8
	macaroni2	78.5±1.3	81.1±0.6	74.1±1.0	97.5±0.2	10.6±0.7	20.0±1.1	80.8±1.1
	pcb1	91.0±1.8	88.6±1.8	86.6±0.9	98.7±0.2	46.0±11.8	51.8±10.0	93.5±0.4
	pcb2	84.3±1.6	82.6±1.0	78.4±0.7	95.8±0.1	18.3±0.9	31.0±1.0	82.8±0.1
	pcb3	87.4±1.8	89.3±1.7	80.7±2.5	95.5±0.2	34.4±1.5	44.1±1.8	85.8±0.2
	pcb4	96.1±2.7	95.7±2.6	91.6±2.7	97.1±0.1	29.3±4.7	36.5±3.9	90.1±0.2
	pipe_fryum	99.6±0.3	99.6±0.2	98.5±0.5	98.5±0.0	37.5±0.4	49.1±0.5	97.8±0.1
	Mean	92.3±0.5	93.0±0.4	88.1±0.5	97.8±0.0	35.8±1.0	44.0±1.0	91.6±0.2
4-shot	candle	94.4±0.8	93.8±1.1	89.6±0.9	99.3±0.1	29.1±1.1	42.5±0.7	97.1±0.2
	capsules	95.5±0.4	97.4±0.3	91.6±0.5	98.0±0.2	41.3±1.2	46.7±1.1	94.6±0.6
	cashew	92.9±0.5	96.6±0.3	89.9±0.7	98.1±0.1	57.0±0.4	61.3±0.4	94.3±0.9
	chewinggum	97.6±0.2	99.0±0.1	96.6±0.3	99.6±0.1	70.2±1.9	71.0±0.2	95.1±0.5
	fryum	98.6±0.1	99.3±0.1	96.5±0.5	96.7±0.2	37.3±1.0	45.6±1.4	93.7±0.2
	macaroni1	93.7±0.4	94.4±0.3	86.3±1.1	99.5±0.1	24.4±5.4	31.2±2.0	94.6±1.0
	macaroni2	80.6±0.8	82.5±0.2	75.5±1.5	97.8±0.1	10.4±1.2	19.3±2.7	81.5±0.7
	pcb1	91.7±0.6	89.2±0.7	87.0±0.6	98.7±0.2	46.7±10.5	52.7±9.1	93.9±0.1
	pcb2	87.8±0.6	85.7±0.9	81.4±0.8	96.0±0.0	19.6±0.7	33.8±0.4	83.7±0.1
	pcb3	89.3±0.6	91.0±1.0	81.3±1.1	95.8±0.1	35.6±1.8	46.5±2.2	86.4±0.3
	pcb4	97.1±1.5	96.7±1.7	92.6±2.1	97.3±0.1	30.0±4.6	37.5±3.9	90.7±0.3
	pipe_fryum	99.8±0.1	99.8±0.1	99.0±0.5	98.6±0.0	38.5±0.2	50.2±0.5	97.8±0.1
	Mean	93.3±0.2	93.8±0.1	88.9±0.3	98.0±0.1	36.7±1.3	44.9±0.9	92.0±0.1

Table A10: Image anomaly classification and pixel anomaly segmentation results on **VisA** with zero-shot and few-shot **UniADet**[‡].

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	**I-AUPR**	**I-F1$_{max}$**	**P-AUROC**	**P-AUPR**	**P-F1$_{max}$**	**P-AUPRO**
0-shot	candle	94.7	95.7	88.9	99.5	51.8	54.8	93.0
	capsules	98.7	99.2	96.4	99.6	73.3	68.9	98.8
	cashew	91.3	95.5	89.8	90.8	16.9	19.6	92.9
	chewinggum	99.4	99.6	97.5	99.7	87.3	79.3	95.3
	fryum	97.2	98.6	94.7	94.2	26.1	30.7	93.8
	macaroni1	98.5	98.3	95.4	99.8	35.5	40.8	98.4
	macaroni2	80.4	80.6	77.8	99.6	16.8	25.9	97.1
	pcb1	92.6	91.7	85.8	91.3	7.8	11.0	89.7
	pcb2	71.7	66.8	71.2	94.3	23.7	31.7	83.5
	pcb3	78.1	73.3	77.4	90.3	4.7	11.3	79.7
	pcb4	93.4	93.1	85.7	96.1	29.4	33.8	88.4
	pipe_fryum	99.3	99.5	97.5	95.2	19.6	28.8	98.3
	Mean	91.3	91.0	88.2	95.9	32.7	36.4	92.4
1-shot	candle	94.9±0.2	94.8±0.3	88.8±0.9	99.6±0.0	50.2±0.5	51.4±0.3	96.4±0.3
	capsules	97.9±0.2	98.7±0.1	94.5±1.2	98.4±0.1	59.1±1.5	61.4±1.6	96.5±0.3
	cashew	96.4±0.7	98.1±0.4	93.3±0.7	97.3±0.6	47.0±1.8	50.3±2.0	95.2±2.1
	chewinggum	98.9±0.1	99.4±0.0	96.9±0.2	99.4±0.1	77.2±0.7	75.3±0.8	94.5±0.5
	fryum	97.2±0.3	98.7±0.1	93.9±0.9	94.3±0.4	32.3±1.2	36.4±1.5	92.1±0.7
	macaroni1	99.0±0.1	98.7±0.1	95.9±0.5	99.6±0.1	20.9±0.4	31.0±1.7	94.9±0.7
	macaroni2	84.2±3.4	85.9±2.9	78.2±2.5	99.1±0.4	12.2±5.2	22.4±6.3	94.0±1.3
	pcb1	91.1±9.0	91.3±7.0	87.3±7.8	98.7±0.4	51.0±0.7	56.7±2.2	90.3±5.7
	pcb2	90.7±1.0	87.7±1.2	86.9±2.7	97.8±0.1	35.0±1.1	44.9±1.3	88.8±0.9
	pcb3	94.8±0.7	95.8±0.6	88.6±0.9	97.5±0.1	45.8±2.3	51.1±1.3	88.4±0.6
	pcb4	98.1±1.0	98.0±0.7	93.6±2.0	97.8±0.5	37.1±1.9	40.0±1.5	90.4±3.6
	pipe_fryum	99.7±0.2	99.7±0.1	98.8±0.3	98.5±0.2	37.9±2.2	50.5±2.1	97.8±0.2
	Mean	95.2±0.8	95.6±0.6	91.4±0.5	98.2±0.1	42.1±0.2	47.6±0.0	93.3±0.5
2-shot	candle	94.6±0.4	94.5±0.3	88.2±1.1	99.6±0.0	50.3±0.5	50.8±0.6	96.6±0.3
	capsules	97.7±1.2	98.7±0.6	94.5±1.4	98.8±0.1	60.8±1.3	61.9±1.2	96.3±1.9
	cashew	96.6±1.0	98.1±0.5	93.5±1.0	97.5±0.4	47.7±0.8	50.6±0.9	95.3±0.8
	chewinggum	99.2±0.2	99.5±0.1	97.3±0.6	99.3±0.1	76.5±1.1	75.2±1.1	94.0±0.5
	fryum	97.4±0.4	98.7±0.2	94.0±1.0	95.0±0.4	34.4±0.9	38.7±1.1	92.6±1.1
	macaroni1	98.5±0.3	98.5±0.3	95.9±0.2	99.6±0.1	22.6±1.6	33.9±2.2	95.0±0.8
	macaroni2	88.9±6.4	90.2±6.3	83.7±4.8	99.4±0.3	16.5±6.8	27.1±8.2	95.6±1.6
	pcb1	95.4±1.6	93.7±2.5	91.7±0.6	99.1±0.2	59.5±9.9	64.3±7.7	93.5±0.1
	pcb2	91.3±0.3	88.3±0.5	87.8±1.1	98.1±0.1	36.0±0.6	45.9±0.3	90.1±0.3
	pcb3	95.6±0.2	96.4±0.2	89.4±0.8	97.9±0.1	46.9±0.2	52.6±0.5	89.6±0.6
	pcb4	99.1±0.0	98.9±0.0	95.5±0.9	98.4±0.1	41.1±1.4	44.1±2.0	93.0±0.9
	pipe_fryum	99.6±0.2	99.6±0.1	98.3±0.8	98.6±0.1	38.6±1.6	50.0±1.5	97.8±0.1
	Mean	96.1±0.8	96.3±0.8	92.5±0.4	98.4±0.0	44.2±0.3	49.6±0.2	94.1±0.3
4-shot	candle	95.3±0.2	95.0±0.4	89.3±0.2	99.6±0.0	50.9±0.7	50.9±0.9	96.7±0.2
	capsules	98.6±0.6	99.1±0.3	95.5±1.8	98.9±0.1	60.9±1.2	61.3±1.2	97.2±1.0
	cashew	96.6±0.8	98.1±0.5	93.1±0.9	97.5±0.4	48.8±1.2	51.8±1.3	95.3±0.7
	chewinggum	99.4±0.2	99.6±0.1	97.4±0.5	99.4±0.1	76.8±1.5	74.7±0.6	93.6±0.9
	fryum	97.6±0.5	98.8±0.2	94.6±0.9	95.5±0.3	36.1±0.7	40.4±0.5	93.3±0.6
	macaroni1	98.7±0.1	98.6±0.2	96.1±0.4	99.7±0.1	23.3±0.9	34.4±1.3	94.7±0.3
	macaroni2	94.6±1.3	94.8±0.8	87.8±1.8	99.7±0.0	20.4±2.0	31.9±2.1	97.4±0.3
	pcb1	95.3±1.9	93.1±2.6	92.2±0.5	99.1±0.2	59.6±9.6	64.1±7.6	93.8±0.2
	pcb2	91.7±0.7	88.7±0.5	89.1±1.0	98.1±0.2	36.2±0.5	46.2±0.4	89.9±0.5
	pcb3	96.5±0.7	97.2±0.4	91.5±1.2	98.1±0.1	46.9±0.9	53.2±0.2	89.8±0.2
	pcb4	99.2±0.1	99.0±0.1	96.0±0.6	98.6±0.1	42.7±1.7	45.5±1.9	94.1±0.4
	pipe_fryum	99.6±0.1	99.6±0.1	99.3±0.3	98.8±0.1	40.1±1.1	51.7±1.3	97.9±0.1
	Mean	96.9±0.3	96.8±0.3	93.5±0.1	98.6±0.1	45.2±0.6	50.5±0.3	94.5±0.2

Table A11: Image anomaly classification and pixel anomaly segmentation results on **VisA** with zero-shot and few-shot **UniADet***.

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	**I-AUPR**	**I-F1$_{max}$**	**P-AUROC**	**P-AUPR**	**P-F1$_{max}$**	**P-AUPRO**
0-shot	candle	94.9	95.8	89.8	99.4	52.6	54.0	92.9
	capsules	98.7	99.4	98.0	99.6	76.8	72.0	99.3
	cashew	95.1	97.6	93.1	93.4	20.1	22.9	95.4
	chewinggum	98.9	99.4	96.9	99.6	88.2	81.3	94.5
	fryum	95.6	97.9	94.0	95.1	30.1	35.5	92.9
	macaroni1	97.3	97.5	92.3	99.7	32.3	40.2	97.9
	macaroni2	80.7	81.6	76.3	99.5	13.4	22.2	97.0
	pcb1	86.1	87.6	79.4	93.1	6.5	11.7	86.9
	pcb2	84.3	83.9	77.9	94.4	19.5	25.7	81.9
	pcb3	77.4	78.6	73.2	90.9	4.7	9.1	77.0
	pcb4	94.6	94.6	88.2	95.8	25.4	29.1	85.8
	pipe_fryum	99.6	99.7	98.0	95.5	20.8	28.1	98.4
	Mean	91.9	92.8	88.1	96.3	32.5	36.0	91.7
1-shot	candle	96.0±0.6	95.8±0.3	90.5±1.2	99.6±0.1	50.6±0.8	54.5±0.3	96.6±0.4
	capsules	99.6±0.3	99.6±0.2	98.2±1.0	99.0±0.2	63.8±1.3	63.0±1.4	98.1±0.3
	cashew	98.5±0.6	99.1±0.3	95.8±1.1	96.9±0.7	46.9±1.4	50.8±1.2	96.1±1.3
	chewinggum	98.7±0.2	99.3±0.1	96.3±0.3	99.2±0.1	81.4±0.6	76.6±0.1	93.6±0.2
	fryum	98.0±0.1	99.1±0.1	96.5±0.5	95.8±0.2	39.4±0.5	42.9±0.2	93.0±0.2
	macaroni1	99.0±0.4	98.8±0.5	96.6±0.8	99.7±0.1	26.8±0.7	35.5±1.1	97.8±0.2
	macaroni2	87.4±1.0	88.0±1.3	80.5±0.9	99.4±0.2	15.2±1.3	25.1±1.7	96.0±0.2
	pcb1	87.2±11.2	85.1±7.6	84.3±10.1	98.3±0.4	45.0±1.8	50.7±4.5	90.5±5.4
	pcb2	92.9±0.6	88.4±0.8	87.5±0.9	97.4±0.1	38.2±1.2	49.9±0.7	86.9±0.6
	pcb3	94.3±0.3	95.4±0.4	89.1±0.2	96.3±0.1	43.4±2.4	49.8±1.7	87.2±0.7
	pcb4	98.3±0.4	98.1±0.2	94.8±1.9	97.6±0.2	37.2±0.8	40.6±0.9	89.7±1.9
	pipe_fryum	100.0±0.1	99.8±0.0	99.5±0.5	98.7±0.2	39.4±2.0	51.7±1.8	98.4±0.1
	Mean	95.8±0.8	95.5±0.5	92.5±0.8	98.2±0.1	44.0±0.1	49.3±0.3	93.6±0.6
2-shot	candle	96.1±0.2	96.0±0.1	90.6±1.2	99.6±0.0	51.5±1.2	54.6±0.6	96.8±0.4
	capsules	99.7±0.2	99.7±0.1	98.7±0.8	99.3±0.1	66.4±1.0	64.8±1.2	98.4±0.4
	cashew	98.5±0.3	99.2±0.2	95.7±0.5	97.0±0.2	47.1±0.6	50.6±0.4	95.7±0.9
	chewinggum	98.8±0.1	99.3±0.1	96.8±0.3	99.1±0.1	81.2±0.4	76.9±0.2	92.8±1.1
	fryum	98.1±0.2	99.2±0.1	96.5±0.5	96.0±0.3	39.5±0.4	42.7±0.3	93.4±0.2
	macaroni1	98.9±0.1	98.9±0.1	95.9±0.5	99.7±0.1	28.0±0.6	37.4±1.2	97.8±0.4
	macaroni2	91.2±3.3	91.6±3.4	85.4±3.5	99.5±0.2	17.5±3.6	28.4±4.9	96.8±0.6
	pcb1	93.5±1.4	88.8±1.7	90.1±0.8	98.7±0.2	50.7±7.4	57.5±6.7	93.1±0.9
	pcb2	93.2±0.3	88.7±0.5	88.5±0.8	97.6±0.1	37.9±0.4	50.1±0.1	87.8±0.8
	pcb3	94.3±0.6	95.6±0.3	90.1±0.8	96.5±0.3	43.5±0.5	50.9±0.9	86.9±0.9
	pcb4	99.1±0.2	98.9±0.2	96.3±0.6	97.8±0.1	39.1±1.8	43.0±2.3	90.8±0.7
	pipe_fryum	99.9±0.1	99.8±0.0	99.3±0.3	98.8±0.1	40.4±1.0	52.3±1.0	98.5±0.1
	Mean	96.8±0.4	96.3±0.4	93.7±0.4	98.3±0.0	45.3±0.4	50.8±0.3	94.0±0.4
4-shot	candle	96.9±0.3	96.9±0.4	91.9±0.4	99.6±0.0	51.7±1.0	55.1±0.4	96.9±0.3
	capsules	99.8±0.1	99.8±0.1	98.8±0.6	99.5±0.1	67.3±0.8	65.2±1.0	98.6±0.3
	cashew	98.3±0.4	99.0±0.3	95.2±0.8	97.0±0.3	47.7±0.5	51.0±0.3	95.9±0.5
	chewinggum	98.9±0.1	99.4±0.1	97.1±0.3	99.1±0.1	81.3±0.4	76.9±0.1	92.4±1.0
	fryum	98.1±0.2	99.2±0.1	96.0±0.1	96.2±0.2	39.5±0.4	42.7±0.2	93.6±0.2
	macaroni1	98.9±0.1	99.0±0.1	96.0±1.3	99.8±0.0	29.0±0.5	39.0±0.3	97.7±0.1
	macaroni2	93.4±0.7	93.9±0.9	88.3±2.5	99.6±0.1	19.0±1.0	30.5±2.4	97.5±0.1
	pcb1	93.5±1.1	88.5±1.3	90.4±1.0	98.7±0.3	49.8±6.6	56.8±6.2	93.2±0.6
	pcb2	93.5±0.8	88.8±0.9	88.5±0.4	97.7±0.1	37.7±1.1	50.4±0.3	87.8±0.2
	pcb3	93.9±0.5	95.6±0.3	90.3±0.5	96.6±0.2	42.7±0.5	51.7±1.1	86.7±0.4
	pcb4	99.2±0.3	99.0±0.4	96.6±0.6	97.9±0.1	39.1±2.0	43.3±2.6	90.9±0.2
	pipe_fryum	100.0±0.0	99.8±0.0	99.8±0.3	98.9±0.0	41.0±0.2	53.3±0.4	98.5±0.1
	Mean	97.1±0.1	96.6±0.1	94.1±0.3	98.4±0.1	45.5±0.6	51.3±0.4	94.1±0.2

Table A12: Image anomaly classification and pixel anomaly segmentation results on **BTAD** with zero-shot and few-shot **UniADet**†.

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
0-shot	1	98.1	99.2	97.0	95.2	52.8	57.4	82.6
	2	91.6	98.7	95.0	94.9	61.3	61.4	69.2
	3	99.6	95.6	89.4	97.1	14.4	25.7	91.9
	Mean	96.4	97.8	93.8	95.7	42.8	48.2	81.3
1-shot	1	98.4±0.2	99.4±0.0	98.0±0.0	97.3±0.1	63.7±0.2	64.4±0.1	85.6±0.6
	2	92.1±1.2	98.7±0.2	95.6±0.2	93.1±0.2	53.2±0.6	53.6±0.6	67.7±1.1
	3	99.3±0.3	94.0±1.2	87.6±1.6	98.2±0.4	46.2±7.6	49.7±5.0	94.4±1.3
	Mean	96.6±0.3	97.4±0.4	93.7±0.5	96.2±0.1	54.4±2.5	55.9±1.6	82.6±0.4
2-shot	1	98.3±0.2	99.4±0.1	98.0±0.0	97.4±0.1	63.8±0.2	64.5±0.1	85.3±0.4
	2	92.0±0.5	98.7±0.1	95.7±0.4	93.3±0.1	53.3±0.4	53.8±0.2	66.7±2.2
	3	99.3±0.3	93.9±1.2	87.6±2.0	98.5±0.2	46.6±6.2	49.6±3.7	95.4±0.5
	Mean	96.5±0.2	97.3±0.4	93.8±0.8	96.4±0.1	54.5±2.1	55.9±1.3	82.5±0.8
4-shot	1	98.4±0.2	99.4±0.1	99.0±0.0	97.5±0.0	64.1±0.3	64.6±0.1	85.3±0.6
	2	92.0±0.5	98.7±0.1	95.5±0.3	93.5±0.1	53.6±0.4	54.2±0.3	65.6±0.4
	3	99.6±0.2	95.3±1.2	89.8±2.7	98.7±0.1	51.7±4.4	52.6±2.7	95.8±0.1
	Mean	96.7±0.2	97.8±0.4	94.8±1.0	96.6±0.1	56.5±1.4	57.1±1.0	82.3±0.2

Table A13: Image anomaly classification and pixel anomaly segmentation results on **BTAD** with zero-shot and few-shot **UniADet**‡.

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
0-shot	1	94.9	98.3	94.6	97.4	65.2	63.7	85.1
	2	90.1	98.4	95.2	96.9	74.6	68.9	75.7
	3	98.1	92.4	89.2	99.0	42.0	45.0	97.4
	Mean	94.4	96.3	93.0	97.7	60.6	59.2	86.1
1-shot	1	98.7±0.3	99.3±0.1	97.2±1.2	98.2±0.1	68.7±0.9	66.2±0.7	86.0±0.5
	2	90.9±0.6	98.5±0.1	94.6±0.3	97.1±0.0	74.0±0.1	67.8±0.2	66.3±1.0
	3	99.5±0.2	95.3±1.3	89.7±3.4	99.7±0.0	75.2±1.8	71.9±1.7	99.0±0.0
	Mean	96.4±0.3	97.7±0.4	93.8±1.3	98.4±0.0	72.7±0.7	68.6±0.6	83.7±0.5
2-shot	1	99.0±0.4	99.4±0.1	97.9±0.0	98.3±0.0	69.0±0.4	66.4±0.3	85.7±0.6
	2	91.7±1.0	98.6±0.2	95.5±0.6	97.2±0.1	74.0±0.1	67.5±0.1	66.5±0.9
	3	99.6±0.1	96.5±1.0	91.9±2.0	99.8±0.1	76.4±1.4	72.6±1.1	99.1±0.1
	Mean	96.8±0.4	98.2±0.4	95.1±0.7	98.4±0.0	73.1±0.5	68.8±0.4	83.8±0.4
4-shot	1	99.1±0.7	99.4±0.2	97.9±0.1	98.3±0.1	69.2±0.7	66.6±0.6	86.2±1.2
	2	91.8±1.5	98.6±0.3	95.2±0.3	97.2±0.1	74.1±0.2	67.2±0.2	65.7±0.3
	3	99.8±0.1	97.2±0.9	92.6±3.4	99.8±0.0	78.5±1.5	73.7±1.2	99.3±0.0
	Mean	96.9±0.3	98.4±0.3	95.3±1.1	98.4±0.1	73.9±0.5	69.2±0.4	83.7±0.3

Table A14: Image anomaly classification and pixel anomaly segmentation results on **BTAD** with zero-shot and few-shot **UniADet***.

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
0-shot	1	99.9	99.7	99.0	97.9	68.6	64.8	90.0
	2	85.0	97.5	93.4	96.8	75.4	69.6	77.4
	3	99.2	93.8	86.8	98.7	32.5	39.1	96.0
	Mean	94.7	97.0	93.1	97.8	58.9	57.8	87.8
1-shot	1	100.0±0.1	99.7±0.0	99.7±0.6	98.0±0.1	68.5±0.6	66.1±0.7	88.8±0.5
	2	91.7±1.6	98.6±0.3	95.2±0.2	97.3±0.1	73.7±0.3	66.6±0.2	70.7±0.8
	3	99.8±0.1	97.2±0.5	93.5±0.6	99.7±0.1	73.0±3.9	67.2±3.2	98.8±0.0
	Mean	97.1±0.5	98.5±0.2	96.1±0.3	98.3±0.1	71.7±1.0	66.7±0.8	86.1±0.4
2-shot	1	100.0±0.0	99.7±0.0	100.0±0.0	98.1±0.1	69.2±0.2	66.7±0.4	89.3±0.5
	2	92.1±1.6	98.7±0.3	95.4±0.2	97.3±0.0	73.4±0.4	66.4±0.3	70.3±1.5
	3	99.8±0.0	97.5±0.3	94.3±0.6	99.7±0.0	73.0±3.6	67.0±2.8	98.9±0.1
	Mean	97.3±0.6	98.6±0.2	96.5±0.2	98.4±0.0	71.8±1.3	66.7±0.9	86.2±0.7
4-shot	1	100.0±0.0	99.7±0.0	100.0±0.0	98.2±0.1	69.2±0.5	66.7±0.5	89.6±0.5
	2	92.6±0.8	98.8±0.1	95.4±0.3	97.3±0.0	73.4±0.2	66.4±0.2	69.9±0.7
	3	99.8±0.1	97.9±0.5	95.4±1.7	99.8±0.1	75.5±3.3	69.0±3.0	99.1±0.1
	Mean	97.5±0.3	98.8±0.2	96.9±0.6	98.4±0.0	72.7±1.0	67.4±0.9	86.2±0.4

Table A15: Image anomaly classification and pixel anomaly segmentation results on **DTD** with zero-shot and few-shot **UniADet**†.

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
0-shot	Woven_001	100.0	99.8	100.0	99.8	75.8	69.6	99.5
	Woven_127	95.9	97.1	92.7	95.1	54.1	55.3	92.4
	Woven_104	99.1	99.7	98.8	97.6	69.3	64.4	95.8
	Stratified_154	99.1	99.8	99.4	99.2	68.4	65.7	99.2
	Blotchy_099	100.0	99.9	100.0	98.8	59.3	56.5	95.0
	Woven_068	96.9	98.3	94.0	98.9	50.9	48.1	96.8
	Woven_125	100.0	99.9	100.0	99.6	76.6	69.6	98.8
	Marbled_078	99.1	99.7	98.1	99.0	62.1	61.4	97.1
	Perforated_037	99.6	99.8	98.8	97.7	67.4	63.8	95.6
	Mesh_114	89.8	96.0	88.4	96.6	63.6	62.5	87.1
	Fibrous_183	99.3	99.7	98.1	99.1	67.1	60.9	97.7
	Matted_069	98.5	99.5	96.9	99.0	57.3	56.1	88.0
	Mean	98.1	99.1	97.1	98.4	64.3	61.2	95.2
1-shot	Woven_001	100.0±0.1	99.8±0.0	99.3±0.0	99.6±0.0	77.9±0.3	73.6±0.2	98.8±0.1
	Woven_127	97.2±0.6	98.1±0.2	95.2±0.4	92.7±0.3	59.8±0.6	63.2±0.7	89.6±0.7
	Woven_104	96.5±0.1	99.1±0.1	95.3±0.3	94.0±0.1	60.7±1.8	60.8±1.2	90.1±0.4
	Stratified_154	99.7±0.4	99.8±0.1	99.0±1.0	99.3±0.1	74.5±0.9	69.9±0.8	98.7±0.2
	Blotchy_099	100.0±0.0	99.9±0.0	100.0±0.0	99.7±0.0	83.3±0.5	77.7±0.7	97.6±0.2
	Woven_068	96.3±1.7	98.1±0.9	94.1±2.8	98.0±0.1	68.2±1.3	64.4±1.3	94.4±0.2
	Woven_125	100.0±0.0	99.9±0.0	100.0±0.0	99.5±0.1	84.3±0.2	78.2±0.1	98.7±0.0
	Marbled_078	100.0±0.0	99.9±0.0	100.0±0.0	99.3±0.0	80.1±0.4	76.0±0.4	98.2±0.0
	Perforated_037	96.8±1.2	99.1±0.3	97.5±0.6	93.4±0.5	63.5±1.7	63.3±1.4	89.4±1.0
	Mesh_114	90.4±1.2	96.3±0.4	88.2±1.0	94.1±0.3	62.4±0.3	62.0±0.6	81.7±0.5
	Fibrous_183	99.8±0.2	99.8±0.1	99.0±0.9	99.5±0.0	86.0±0.1	80.1±0.2	98.8±0.1
	Matted_069	99.1±0.7	99.7±0.2	99.0±0.3	99.6±0.1	79.9±7.3	75.1±4.7	92.2±1.3
	Mean	98.0±0.2	99.1±0.1	97.2±0.4	97.4±0.1	73.4±0.9	70.3±0.7	94.0±0.2
2-shot	Woven_001	99.9±0.1	99.8±0.0	99.5±0.4	99.7±0.1	78.3±0.3	73.6±0.2	98.9±0.1
	Woven_127	96.9±0.3	98.0±0.1	94.9±0.4	93.5±0.3	61.9±0.7	64.0±0.2	90.9±0.6
	Woven_104	96.8±0.2	99.2±0.1	95.1±0.3	94.5±0.3	62.0±1.8	61.4±0.9	90.9±0.6
	Stratified_154	99.8±0.1	99.8±0.1	99.0±0.3	99.4±0.1	74.6±0.4	69.8±0.4	98.8±0.1
	Blotchy_099	100.0±0.0	99.9±0.0	100.0±0.0	99.7±0.0	84.4±0.9	78.2±0.7	97.7±0.1
	Woven_068	95.9±2.3	97.9±1.2	94.2±3.4	98.1±0.1	68.6±1.4	64.6±1.4	94.4±0.2
	Woven_125	100.0±0.0	99.9±0.0	100.0±0.0	99.6±0.0	84.5±0.2	78.2±0.2	98.8±0.0
	Marbled_078	100.0±0.0	99.9±0.0	100.0±0.0	99.3±0.0	80.7±0.4	75.9±0.3	98.3±0.0
	Perforated_037	96.6±0.9	99.0±0.2	97.5±0.6	93.6±0.1	63.9±0.4	63.1±0.4	90.1±0.0
	Mesh_114	91.0±1.2	96.6±0.4	90.4±1.7	94.4±0.2	62.7±0.6	62.3±0.1	82.2±0.5
	Fibrous_183	99.7±0.3	99.8±0.1	98.7±1.1	99.6±0.1	86.7±0.4	80.3±0.2	98.9±0.1
	Matted_069	99.4±0.7	99.7±0.2	98.5±0.8	99.7±0.1	83.7±6.7	77.1±4.3	92.1±1.2
	Mean	98.0±0.1	99.1±0.1	97.3±0.3	97.6±0.0	74.3±0.7	70.7±0.5	94.3±0.1
4-shot	Woven_001	99.9±0.1	99.8±0.0	99.5±0.4	99.7±0.1	78.4±0.3	73.5±0.4	98.9±0.1
	Woven_127	97.6±0.4	98.4±0.3	96.3±1.4	93.8±0.5	63.5±1.0	65.5±1.3	91.6±0.8
	Woven_104	98.1±1.5	99.4±0.3	97.0±2.1	94.6±0.1	62.8±1.3	62.2±0.8	91.3±0.1
	Stratified_154	99.9±0.1	99.9±0.0	99.6±0.3	99.4±0.1	74.0±1.5	68.9±1.7	98.8±0.0
	Blotchy_099	100.0±0.0	99.9±0.0	100.0±0.0	99.7±0.0	84.6±0.3	78.4±0.3	97.7±0.1
	Woven_068	95.5±3.1	97.7±1.7	94.1±4.2	98.3±0.1	69.2±2.3	65.1±2.4	95.1±0.3
	Woven_125	100.0±0.0	99.9±0.0	100.0±0.0	99.6±0.0	85.2±0.2	78.5±0.1	98.8±0.0
	Marbled_078	100.0±0.0	99.9±0.0	100.0±0.0	99.3±0.0	81.0±0.2	75.9±0.0	98.3±0.0
	Perforated_037	98.3±1.5	99.4±0.4	98.1±0.7	94.3±0.1	65.4±0.6	64.3±0.7	91.2±0.3
	Mesh_114	91.3±2.1	96.6±0.6	89.4±1.0	94.7±0.1	62.8±0.2	62.3±0.2	83.1±0.3
	Fibrous_183	99.7±0.3	99.8±0.1	98.9±1.0	99.6±0.0	87.4±0.1	80.7±0.3	98.9±0.0
	Matted_069	99.8±0.1	99.8±0.1	99.0±0.4	99.8±0.1	87.8±0.6	79.9±0.7	90.4±1.4
	Mean	98.4±0.4	99.2±0.2	97.7±0.4	97.7±0.1	75.2±0.3	71.3±0.3	94.5±0.1

Table A16: Image anomaly classification and pixel anomaly segmentation results on **DTD** with zero-shot and few-shot **UniADet**[‡].

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
0-shot	Woven_001	100.0	99.8	100.0	99.9	84.5	75.4	99.6
	Woven_127	95.7	96.8	91.9	95.8	69.4	66.8	95.7
	Woven_104	99.8	99.8	99.4	99.4	78.7	71.6	98.7
	Stratified_154	97.4	99.2	96.4	99.6	79.6	74.8	99.3
	Blotchy_099	76.9	94.0	88.9	95.5	54.1	53.4	76.3
	Woven_068	98.7	99.0	95.1	99.4	73.0	67.8	98.4
	Woven_125	99.2	99.7	98.1	99.8	86.9	79.6	99.4
	Marbled_078	90.8	97.8	91.8	96.5	64.8	60.8	90.0
	Perforated_037	100.0	99.9	100.0	99.4	74.6	70.8	98.8
	Mesh_114	95.0	98.0	92.9	98.5	77.3	73.1	95.8
	Fibrous_183	93.1	98.3	92.4	98.9	69.2	62.4	96.2
	Matted_069	84.7	96.0	89.4	97.1	51.2	51.7	80.3
	Mean	94.3	98.2	94.7	98.3	71.9	67.4	94.0
1-shot	Woven_001	100.0±0.0	99.8±0.0	100.0±0.0	99.8±0.1	86.2±0.5	78.5±0.2	99.4±0.2
	Woven_127	98.4±0.2	99.0±0.1	98.3±0.7	95.2±0.1	71.0±0.6	69.5±0.7	94.3±0.1
	Woven_104	100.0±0.0	99.9±0.0	100.0±0.0	99.3±0.1	82.9±0.4	77.5±0.2	98.8±0.1
	Stratified_154	100.0±0.0	99.9±0.0	100.0±0.0	99.6±0.1	79.2±0.7	73.5±0.2	99.2±0.1
	Blotchy_099	97.1±0.3	99.3±0.1	96.2±0.0	99.7±0.1	89.4±0.2	83.4±0.2	95.4±1.0
	Woven_068	99.6±0.6	99.6±0.2	98.6±1.0	98.2±0.2	78.2±1.1	73.6±2.2	95.7±0.3
	Woven_125	100.0±0.1	99.9±0.0	99.8±0.3	99.8±0.0	88.5±0.3	81.6±0.1	99.3±0.1
	Marbled_078	97.5±0.3	99.3±0.1	96.9±0.1	99.7±0.0	84.6±0.3	79.9±0.6	99.0±0.1
	Perforated_037	100.0±0.0	99.9±0.0	100.0±0.0	99.4±0.1	78.2±1.2	73.9±1.0	98.9±0.2
	Mesh_114	95.5±1.3	98.2±0.5	92.8±1.4	97.2±0.5	75.5±1.5	72.2±0.8	92.1±1.2
	Fibrous_183	100.0±0.1	99.9±0.0	99.8±0.3	99.8±0.0	89.0±0.3	81.9±0.6	99.4±0.0
	Matted_069	97.2±0.6	99.2±0.2	96.6±1.0	99.5±0.1	85.4±2.4	77.1±3.8	89.8±2.8
	Mean	98.8±0.1	99.5±0.0	98.2±0.3	98.9±0.1	82.4±0.3	76.9±0.5	96.8±0.2
2-shot	Woven_001	100.0±0.0	99.8±0.0	100.0±0.0	99.9±0.1	86.6±0.2	78.6±0.1	99.5±0.1
	Woven_127	98.3±0.3	99.0±0.2	98.3±0.3	95.3±0.2	71.8±0.6	69.7±0.2	94.9±0.6
	Woven_104	100.0±0.0	99.9±0.0	100.0±0.0	99.4±0.1	83.6±0.2	78.0±0.4	98.9±0.1
	Stratified_154	100.0±0.0	99.9±0.0	100.0±0.0	99.7±0.1	79.3±0.6	73.5±1.0	99.3±0.1
	Blotchy_099	98.4±0.9	99.5±0.2	96.9±0.7	99.8±0.0	90.1±0.4	83.7±0.2	96.3±1.0
	Woven_068	99.5±0.6	99.6±0.2	98.5±1.0	98.3±0.3	78.9±1.6	73.7±2.3	96.0±0.1
	Woven_125	100.0±0.0	99.9±0.0	100.0±0.0	99.8±0.0	88.9±0.2	81.6±0.2	99.4±0.0
	Marbled_078	98.7±0.3	99.5±0.1	97.8±0.4	99.7±0.0	85.0±0.5	80.0±0.7	99.0±0.2
	Perforated_037	100.0±0.0	99.9±0.0	100.0±0.0	99.5±0.1	78.5±0.5	73.8±0.5	99.0±0.2
	Mesh_114	97.5±0.9	98.9±0.4	94.8±1.6	97.6±0.1	76.8±0.4	72.8±0.7	93.3±0.4
	Fibrous_183	100.0±0.0	99.9±0.0	100.0±0.0	99.8±0.0	89.4±0.3	82.1±0.5	99.5±0.1
	Matted_069	98.6±0.7	99.5±0.2	97.9±1.3	99.6±0.0	87.0±1.9	79.1±2.9	90.5±2.1
	Mean	99.2±0.1	99.6±0.0	98.7±0.2	99.0±0.1	83.0±0.1	77.2±0.2	97.1±0.2
4-shot	Woven_001	100.0±0.0	99.8±0.0	100.0±0.0	99.9±0.0	86.8±0.1	78.7±0.2	99.5±0.1
	Woven_127	98.6±0.4	99.1±0.2	98.7±0.0	95.5±0.4	72.5±0.5	69.9±0.3	95.1±0.7
	Woven_104	100.0±0.0	99.9±0.0	100.0±0.0	99.5±0.1	83.8±0.5	78.1±0.4	99.0±0.1
	Stratified_154	100.0±0.0	99.9±0.0	100.0±0.0	99.7±0.0	79.6±1.2	73.5±1.3	99.4±0.1
	Blotchy_099	99.4±0.1	99.8±0.1	98.1±0.0	99.8±0.0	90.5±0.3	83.7±0.2	96.8±0.8
	Woven_068	99.8±0.2	99.7±0.1	98.5±0.3	99.4±0.1	81.7±0.7	74.3±1.2	98.1±0.3
	Woven_125	100.0±0.0	99.9±0.0	100.0±0.0	99.8±0.0	89.1±0.4	81.5±0.4	99.4±0.1
	Marbled_078	99.4±0.2	99.7±0.1	98.8±0.0	99.7±0.1	85.7±0.6	80.3±0.7	99.1±0.2
	Perforated_037	100.0±0.0	99.9±0.0	100.0±0.0	99.5±0.1	79.1±0.5	74.0±0.5	99.1±0.2
	Mesh_114	99.0±0.3	99.5±0.1	97.5±0.0	98.3±0.3	78.5±1.0	73.7±0.7	95.1±0.7
	Fibrous_183	100.0±0.0	99.9±0.0	100.0±0.0	99.8±0.0	89.6±0.4	82.1±0.7	99.4±0.1
	Matted_069	99.7±0.2	99.8±0.0	99.4±0.0	99.7±0.1	88.9±0.9	81.4±0.7	89.2±2.2
	Mean	99.7±0.1	99.7±0.0	99.3±0.1	99.2±0.1	83.8±0.1	77.6±0.1	97.5±0.2

Table A17: Image anomaly classification and pixel anomaly segmentation results on **DTD** with zero-shot and few-shot **UniADet***.

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
0-shot	Woven_001	100.0	99.8	100.0	99.9	85.2	75.6	99.6
	Woven_127	96.5	97.5	92.6	97.8	73.1	68.3	96.6
	Woven_104	99.7	99.8	98.7	99.7	84.7	76.8	99.2
	Stratified_154	97.8	99.5	97.5	99.4	80.1	73.0	99.4
	Blotchy_099	83.1	95.6	88.9	98.3	61.5	57.4	90.5
	Woven_068	100.0	99.8	100.0	98.7	59.7	57.2	97.2
	Woven_125	99.2	99.7	98.1	99.8	87.5	79.2	99.4
	Marbled_078	95.9	99.0	94.3	99.3	81.3	75.1	97.7
	Perforated_037	99.8	99.8	99.4	99.5	76.4	70.4	99.1
	Mesh_114	95.5	98.2	92.2	99.3	81.5	75.3	96.8
	Fibrous_183	100.0	99.9	100.0	99.4	79.4	73.2	98.6
	Matted_069	97.7	99.3	96.3	98.1	60.4	57.2	82.1
	Mean	97.1	99.0	96.5	99.1	75.9	69.9	96.3
1-shot	Woven_001	99.9±0.1	99.8±0.1	99.5±0.8	99.9±0.1	88.6±0.5	80.4±0.3	99.3±0.2
	Woven_127	98.4±0.2	99.0±0.1	98.7±0.0	95.0±0.2	73.9±1.8	71.7±1.2	94.2±0.3
	Woven_104	100.0±0.0	99.9±0.0	100.0±0.0	99.4±0.1	86.9±0.4	79.1±0.4	98.9±0.1
	Stratified_154	100.0±0.0	99.9±0.0	100.0±0.0	99.7±0.1	84.9±0.8	77.0±0.8	99.5±0.1
	Blotchy_099	100.0±0.0	99.9±0.0	100.0±0.0	99.9±0.0	91.5±0.2	84.9±0.3	98.3±0.5
	Woven_068	99.9±0.1	99.8±0.1	99.6±0.8	99.1±0.1	82.6±1.0	76.2±1.7	97.9±0.1
	Woven_125	100.0±0.1	99.9±0.0	99.8±0.3	99.8±0.0	91.1±0.3	82.6±0.1	99.4±0.0
	Marbled_078	100.0±0.0	99.9±0.0	100.0±0.0	99.9±0.0	91.7±0.1	84.5±0.2	99.6±0.0
	Perforated_037	100.0±0.1	99.9±0.0	99.8±0.3	99.3±0.2	79.7±1.1	72.9±0.6	98.9±0.1
	Mesh_114	95.0±0.4	98.1±0.1	93.6±1.4	98.0±0.2	80.5±0.5	74.7±0.4	93.1±0.1
	Fibrous_183	100.0±0.0	99.9±0.0	100.0±0.0	99.7±0.1	89.3±0.5	82.1±0.6	99.2±0.1
	Matted_069	100.0±0.0	99.9±0.0	100.0±0.0	99.7±0.1	89.6±2.0	80.9±3.5	90.0±1.4
	Mean	99.4±0.1	99.6±0.0	99.3±0.1	99.1±0.0	85.8±0.2	78.9±0.4	97.3±0.1
2-shot	Woven_001	100.0±0.0	99.8±0.0	100.0±0.0	99.9±0.0	88.9±0.3	80.5±0.2	99.5±0.1
	Woven_127	98.5±0.2	99.0±0.1	98.7±0.0	95.2±0.4	74.6±0.3	71.6±0.4	94.9±0.4
	Woven_104	100.0±0.0	99.9±0.0	100.0±0.0	99.5±0.2	87.1±0.2	79.4±0.3	99.1±0.2
	Stratified_154	100.0±0.0	99.9±0.0	100.0±0.0	99.8±0.0	83.9±1.2	76.3±1.0	99.6±0.1
	Blotchy_099	100.0±0.0	99.9±0.0	100.0±0.0	99.9±0.0	91.5±0.1	84.9±0.0	98.5±0.3
	Woven_068	99.9±0.1	99.8±0.1	99.6±0.8	99.1±0.1	83.0±1.1	76.5±1.8	98.0±0.2
	Woven_125	100.0±0.0	99.9±0.0	100.0±0.0	99.8±0.0	91.3±0.1	82.7±0.0	99.4±0.1
	Marbled_078	100.0±0.0	99.9±0.0	100.0±0.0	99.9±0.0	91.9±0.1	84.6±0.1	99.6±0.0
	Perforated_037	100.0±0.0	99.9±0.0	100.0±0.0	99.4±0.0	79.4±0.3	72.6±0.2	99.0±0.1
	Mesh_114	96.8±1.4	98.7±0.4	95.0±0.4	98.2±0.1	81.1±0.2	75.6±0.4	93.3±0.1
	Fibrous_183	100.0±0.0	99.9±0.0	100.0±0.0	99.7±0.1	89.7±0.4	82.3±0.4	99.3±0.1
	Matted_069	100.0±0.0	99.9±0.0	100.0±0.0	99.7±0.1	90.9±1.8	82.9±2.9	90.2±0.3
	Mean	99.6±0.1	99.7±0.1	99.5±0.1	99.2±0.1	86.1±0.2	79.2±0.3	97.5±0.0
4-shot	Woven_001	100.0±0.0	99.8±0.0	100.0±0.0	99.9±0.0	88.9±0.2	80.6±0.1	99.6±0.1
	Woven_127	98.5±0.2	99.1±0.1	98.7±0.0	95.5±0.3	75.3±0.5	71.9±0.2	95.2±0.6
	Woven_104	100.0±0.0	99.9±0.0	100.0±0.0	99.6±0.1	87.5±0.3	79.5±0.3	99.1±0.1
	Stratified_154	100.0±0.0	99.9±0.0	100.0±0.0	99.8±0.1	84.3±1.2	76.5±1.2	99.6±0.0
	Blotchy_099	100.0±0.0	99.9±0.0	100.0±0.0	99.9±0.0	91.5±0.1	85.1±0.2	98.7±0.2
	Woven_068	100.0±0.0	99.8±0.0	100.0±0.0	99.4±0.1	85.1±0.5	77.9±0.6	98.6±0.2
	Woven_125	100.0±0.0	99.9±0.0	100.0±0.0	99.8±0.1	91.5±0.3	82.9±0.3	99.5±0.0
	Marbled_078	100.0±0.0	99.9±0.0	100.0±0.0	99.9±0.0	91.9±0.2	84.5±0.3	99.6±0.0
	Perforated_037	100.0±0.0	99.9±0.0	100.0±0.0	99.5±0.1	80.2±0.8	72.9±0.3	99.1±0.1
	Mesh_114	97.5±1.1	99.0±0.4	96.1±1.2	98.5±0.2	82.1±0.3	76.0±0.2	94.6±0.7
	Fibrous_183	100.0±0.0	99.9±0.0	100.0±0.0	99.8±0.1	90.0±0.3	82.5±0.3	99.4±0.1
	Matted_069	100.0±0.0	99.9±0.0	100.0±0.0	99.8±0.1	92.0±0.1	84.7±0.2	88.6±2.2
	Mean	99.7±0.1	99.7±0.1	99.6±0.1	99.3±0.1	86.7±0.1	79.6±0.1	97.7±0.2

Table A18: Image anomaly classification and pixel anomaly segmentation results on **KSDD** with zero-shot and few-shot **UniADet**†.

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
0-shot	electrical_commutators	96.3	83.4	80.0	98.1	31.8	43.6	86.3
1-shot	electrical_commutators	97.7$\pm$0.7	92.4$\pm$3.2	89.7$\pm$2.7	98.0$\pm$0.0	45.0$\pm$5.8	50.9$\pm$3.9	87.0$\pm$0.1
2-shot	electrical_commutators	97.7$\pm$0.6	93.3$\pm$1.3	89.4$\pm$3.6	98.1$\pm$0.2	46.1$\pm$3.9	50.6$\pm$3.6	87.4$\pm$0.6
4-shot	electrical_commutators	97.3$\pm$0.8	91.3$\pm$1.8	87.1$\pm$3.6	98.1$\pm$0.2	42.5$\pm$5.7	47.9$\pm$5.3	87.4$\pm$0.7

Table A19: Image anomaly classification and pixel anomaly segmentation results on **KSDD** with zero-shot and few-shot **UniADet**‡.

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
0-shot	electrical_commutators	91.6	59.5	73.0	98.4	15.8	24.5	90.2
1-shot	electrical_commutators	94.7$\pm$0.4	68.7$\pm$2.0	78.7$\pm$1.5	99.4$\pm$0.3	34.9$\pm$2.7	46.1$\pm$0.9	97.7$\pm$0.6
2-shot	electrical_commutators	95.3$\pm$0.2	71.2$\pm$1.2	78.8$\pm$1.4	99.5$\pm$0.1	36.9$\pm$2.7	47.7$\pm$1.6	97.6$\pm$0.3
4-shot	electrical_commutators	95.3$\pm$0.6	71.7$\pm$1.9	79.2$\pm$1.1	99.4$\pm$0.2	35.2$\pm$5.3	45.6$\pm$5.1	96.8$\pm$1.5

Table A20: Image anomaly classification and pixel anomaly segmentation results on **KSDD** with zero-shot and few-shot **UniADet***.

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
0-shot	electrical_commutators	91.0	64.5	67.8	99.0	22.6	32.6	92.0
1-shot	electrical_commutators	96.9$\pm$0.3	84.7$\pm$2.0	83.4$\pm$0.9	99.3$\pm$0.1	45.2$\pm$3.7	50.4$\pm$2.4	94.7$\pm$0.3
2-shot	electrical_commutators	97.2$\pm$0.4	87.5$\pm$0.9	84.0$\pm$1.3	99.3$\pm$0.2	45.0$\pm$2.2	49.9$\pm$1.2	94.6$\pm$0.4
4-shot	electrical_commutators	96.7$\pm$0.2	86.5$\pm$1.7	82.4$\pm$1.4	99.3$\pm$0.2	44.0$\pm$5.7	49.3$\pm$4.1	93.9$\pm$2.3

Table A21: Image anomaly classification results (I-AUROC, I-AUPR, I-F1$_{max}$) on **Medical** datasets with zero-shot **UniADet**.

Dataset	UniADet†			UniADet†			UniADet*		
HeadCT	97.0	96.7	91.8	95.2	94.9	89.2	95.7	94.8	90.1
BrainMRI	95.7	96.4	95.1	91.7	94.8	88.0	91.9	95.0	88.3
Br35H	97.4	97.1	93.0	92.5	93.9	85.7	92.3	93.8	85.3
Mean	96.7	96.7	93.3	93.1	94.5	87.7	93.3	94.5	87.9

Table A22: Pixel anomaly segmentation results (P-AUROC, P-AUPR, P-F1$_{max}$, P-AUPRO) on **Medical** datasets with zero-shot **UniADet***.

Dataset	UniADet†				UniADet†				UniADet*			
ISIC	92.6	84.2	76.8	87.8	95.0	88.5	81.4	90.4	95.9	90.6	83.6	92.6
ColonDB	84.0	36.2	40.8	79.2	86.0	35.5	43.1	79.8	86.7	33.3	43.2	80.6
ClinicDB	90.1	55.8	55.6	81.1	91.3	59.3	60.6	82.2	91.5	57.5	59.5	83.4
Endo	90.7	62.5	61.5	79.3	92.6	67.3	66.4	82.7	92.8	68.5	66.1	83.3
Kvasir	87.5	57.5	58.6	59.7	91.1	65.2	65.3	53.1	91.7	65.9	65.8	54.9
Mean	89.0	59.2	58.7	77.4	91.2	63.2	63.4	77.6	91.7	63.2	63.6	79.0

Table A23: Image anomaly classification and pixel anomaly segmentation results on **Real-IAD** with zero-shot and 1-shot **UniADet**[†].

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
0-shot	audiojack	67.2	56.7	53.5	97.6	27.7	37.6	84.3
	bottle_cap	64.7	55.0	62.8	99.1	15.7	21.5	95.4
	button_battery	81.5	86.3	78.1	98.0	45.0	46.5	89.4
	end_cap	71.6	78.7	74.5	96.2	22.3	30.3	88.4
	eraser	89.9	89.3	79.5	99.8	44.4	48.7	98.2
	fire_hood	83.3	73.7	69.1	99.3	38.4	44.1	93.1
	mint	70.9	71.4	66.1	96.0	26.6	33.9	86.7
	mounts	84.8	70.0	75.3	99.1	40.6	46.1	97.3
	pcb	68.6	79.0	76.1	96.9	18.8	25.1	89.9
	phone_battery	85.7	85.7	74.4	82.1	35.5	42.1	96.1
	plastic_nut	79.7	69.8	61.8	98.2	39.2	41.7	91.0
	plastic_plug	78.3	71.5	64.3	98.9	27.3	34.8	94.9
	porcelain_doll	92.5	88.7	79.7	99.7	42.3	44.2	97.9
	regulator	65.2	47.8	46.8	95.7	13.8	20.7	77.3
	rolled_strip_base	88.0	93.9	84.9	99.5	41.8	45.4	98.4
	sim_card_set	95.4	96.3	89.2	99.8	67.3	63.6	98.2
	switch	71.2	77.5	70.1	87.4	12.8	19.5	84.1
	tape	91.9	90.9	83.6	99.6	53.2	50.4	96.8
	terminalblock	78.2	82.4	73.3	98.7	41.0	46.3	96.2
	toothbrush	71.0	78.0	70.2	93.8	18.4	24.7	71.0
	toy	66.7	75.7	74.3	85.8	15.9	24.5	74.7
	toy_brick	75.0	70.5	64.2	98.5	36.6	40.9	90.1
	transistor1	74.9	80.6	75.8	95.8	34.8	41.4	84.1
	u_block	74.5	60.6	58.7	99.4	31.1	38.1	96.8
	usb	77.3	77.5	69.1	97.5	25.6	33.3	91.7
	usb_adaptor	71.6	66.5	61.1	98.9	25.4	32.7	93.0
	vcpill	82.4	80.0	68.0	98.6	51.8	51.5	89.1
	wooden_beads	86.3	84.6	75.4	99.1	40.3	44.2	94.7
	woodstick	82.0	68.9	63.9	98.5	58.1	56.9	90.6
	zipper	87.2	92.8	82.9	91.0	17.3	24.5	85.9
	Mean	78.6	76.7	70.9	96.6	33.6	38.5	90.5
1-shot	audiojack	75.4±1.9	69.0±2.8	58.9±2.2	98.4±0.1	28.6±2.4	39.4±2.7	90.0±0.4
	bottle_cap	82.6±0.1	78.3±0.6	76.7±0.6	99.7±0.0	35.8±0.2	39.9±0.3	97.2±0.1
	button_battery	77.7±0.3	84.6±0.2	75.6±0.3	96.9±0.2	41.7±1.2	44.0±0.4	83.8±1.3
	end_cap	74.8±0.4	80.7±0.3	75.8±0.3	98.6±0.1	23.4±1.9	29.8±2.5	94.7±0.3
	eraser	92.0±2.6	90.7±2.7	81.0±4.3	99.8±0.1	40.3±1.0	44.6±0.5	98.4±0.4
	fire_hood	85.6±2.9	76.3±3.9	72.6±3.5	99.4±0.2	31.2±7.6	38.2±5.5	92.5±1.9
	mint	68.0±1.7	68.6±1.8	65.3±0.5	95.0±0.2	19.5±1.0	31.2±0.8	78.4±1.3
	mounts	85.1±0.2	71.1±0.2	75.1±0.6	99.1±0.1	35.5±0.3	40.8±0.2	97.0±0.1
	pcb	82.0±1.2	89.0±1.1	79.3±0.4	98.6±0.1	38.6±2.0	43.2±2.2	93.1±0.1
	phone_battery	85.0±2.4	84.5±2.3	72.9±2.9	95.7±0.2	36.7±2.2	42.8±1.9	94.7±0.3
	plastic_nut	84.1±1.0	76.1±2.9	68.0±1.2	99.2±0.0	38.4±0.5	39.5±0.5	96.1±0.1
	plastic_plug	79.1±1.8	73.9±2.4	65.1±1.2	99.0±0.1	28.3±0.4	36.2±0.3	95.2±0.6
	porcelain_doll	91.0±0.7	85.6±1.4	75.4±0.6	99.6±0.1	38.6±1.5	43.6±1.4	97.7±0.3
	regulator	64.6±1.4	45.6±2.5	47.6±0.6	96.9±0.1	6.1±0.8	10.2±0.9	85.3±0.4
	rolled_strip_base	90.9±1.2	95.4±0.5	87.6±1.1	99.5±0.1	32.7±3.0	43.4±2.8	98.2±0.3
	sim_card_set	97.0±0.1	97.6±0.1	91.9±0.2	99.8±0.0	62.8±0.5	59.6±0.4	97.9±0.2
	switch	82.4±1.0	86.2±0.8	75.5±0.3	96.3±0.1	36.2±0.9	43.1±0.8	93.4±0.3
	tape	93.3±0.4	92.2±0.4	83.8±0.3	99.7±0.0	47.4±1.1	50.1±0.6	97.5±0.1
	terminalblock	88.2±0.9	91.2±0.7	80.8±1.5	99.5±0.1	47.7±0.6	50.8±0.6	97.9±0.2
	toothbrush	77.9±1.8	83.2±1.3	72.9±1.6	94.5±0.8	24.6±3.6	33.3±4.7	76.5±3.7
	toy	64.4±7.9	72.8±5.7	75.0±1.1	90.3±0.6	9.4±1.2	13.6±1.2	80.6±0.5
	toy_brick	73.1±2.8	69.0±3.7	62.5±1.8	96.9±0.2	29.3±4.1	35.9±3.9	83.4±1.7
	transistor1	77.8±3.3	82.5±2.6	77.5±1.3	95.7±0.3	29.7±3.6	37.3±4.7	83.8±1.0
	u_block	86.2±1.0	82.6±1.3	73.7±1.9	99.6±0.1	49.4±2.1	52.6±2.6	97.3±0.4
	usb	85.6±0.5	85.3±0.6	76.7±0.3	98.8±0.1	32.6±0.4	39.9±0.3	95.3±0.2
	usb_adaptor	82.5±0.5	75.9±1.3	71.2±0.4	99.4±0.0	27.9±0.9	34.2±1.0	95.5±0.3
	vcpill	88.3±1.1	86.9±1.1	75.6±1.2	98.8±0.1	55.4±2.6	58.4±2.1	89.0±0.6
	wooden_beads	86.4±1.0	86.2±0.9	75.7±0.8	99.2±0.1	45.0±1.0	49.1±0.2	92.6±0.6
	woodstick	82.8±1.6	71.7±1.2	64.3±1.6	97.8±0.2	54.4±0.9	57.1±0.3	83.9±1.5
	zipper	90.8±2.4	94.4±1.4	86.7±2.1	97.5±0.4	32.1±1.9	35.7±1.9	93.1±0.8
	Mean	82.5±0.6	80.9±0.6	74.0±0.5	98.0±0.0	35.3±0.2	40.6±0.1	91.7±0.3

Table A24: Image anomaly classification and pixel anomaly segmentation results on **Real-IAD** with 2-shot and 4-shot **UniADet**†.

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
2-shot	audiojack	77.1±0.6	71.6±0.5	61.0±0.7	98.5±0.1	30.5±1.2	42.2±0.8	90.8±0.4
	bottle_cap	82.3±0.1	77.6±0.2	77.1±0.3	99.7±0.1	35.7±0.4	39.9±0.2	97.4±0.1
	button_battery	79.0±1.0	85.5±0.8	75.7±0.9	97.4±0.2	48.0±3.4	48.9±3.3	86.3±1.4
	end_cap	76.0±1.4	81.6±1.3	76.1±0.2	98.7±0.0	25.1±0.8	31.9±1.1	95.1±0.2
	eraser	93.5±0.5	92.4±0.3	83.8±0.3	99.8±0.0	40.8±1.5	44.3±0.6	98.7±0.0
	fire_hood	86.0±1.6	76.6±2.6	72.6±2.1	99.5±0.0	30.4±4.9	37.9±3.1	93.2±0.2
	mint	69.3±1.8	70.0±1.9	65.7±0.6	95.3±0.1	19.6±1.2	30.9±1.1	79.3±0.7
	mounts	85.2±0.3	71.3±0.3	75.4±0.6	99.2±0.0	35.1±0.2	40.3±0.2	97.3±0.0
	pcb	83.9±0.6	90.3±0.4	80.4±0.2	98.9±0.1	42.6±0.8	46.5±0.7	94.2±0.3
	phone_battery	87.1±0.6	86.4±0.5	75.4±0.8	96.4±0.1	38.3±0.6	43.6±0.3	95.4±0.1
	plastic_nut	85.1±0.9	77.5±2.3	69.2±0.8	99.3±0.0	38.6±0.3	39.5±0.4	96.6±0.1
	plastic_plug	80.1±1.7	75.1±2.3	65.5±1.4	99.0±0.1	28.4±0.5	36.1±0.7	95.1±0.6
	porcelain_doll	90.8±0.4	84.9±0.8	76.1±1.0	99.6±0.1	36.2±1.8	41.4±1.6	97.8±0.2
	regulator	68.0±1.2	48.9±1.7	50.2±1.0	97.6±0.1	8.0±0.2	14.5±0.7	88.0±0.3
	rolled_strip_base	91.9±1.1	95.8±0.5	88.5±0.9	99.6±0.2	33.5±1.7	44.1±0.7	98.6±0.4
	sim_card_set	96.8±0.1	97.4±0.1	91.7±0.1	99.8±0.0	61.9±0.9	58.7±0.9	97.9±0.2
	switch	83.2±0.7	86.9±0.7	75.9±0.4	96.6±0.1	37.7±0.9	44.7±1.1	94.1±0.2
	tape	93.2±0.5	92.0±0.5	83.5±0.6	99.7±0.0	47.1±1.4	49.9±0.5	97.7±0.1
	terminalblock	88.0±0.5	90.9±0.5	80.9±0.4	99.5±0.1	47.5±0.9	50.5±0.6	98.1±0.2
	toothbrush	77.3±1.6	82.5±1.0	72.6±1.3	94.6±1.0	22.9±4.3	31.7±5.4	77.0±4.2
	toy	72.6±4.4	80.4±2.7	75.6±1.5	90.8±0.7	19.2±1.7	27.3±2.0	82.7±1.1
	toy_brick	73.9±3.5	70.0±4.3	63.2±2.4	97.2±0.3	30.7±3.7	36.8±3.6	83.7±1.2
	transistor1	80.3±0.5	84.6±0.1	78.4±0.3	96.3±0.1	32.1±1.1	39.8±1.8	85.8±0.3
	u_block	86.5±1.0	83.2±1.6	75.2±2.0	99.7±0.1	50.5±2.5	52.7±2.8	97.7±0.4
	usb	86.5±0.4	86.1±0.7	77.7±0.5	99.0±0.1	33.4±0.5	40.1±0.5	96.0±0.2
	usb_adaptor	82.3±0.2	75.5±0.5	71.3±0.2	99.4±0.0	27.9±0.4	33.8±0.5	95.7±0.1
	vcpill	89.6±0.4	88.1±0.5	76.9±0.8	99.1±0.1	57.7±2.6	60.4±1.4	90.5±0.7
	wooden_beads	87.4±0.5	87.0±0.4	77.0±0.4	99.3±0.0	45.1±0.2	48.5±0.1	93.4±0.4
	woodstick	84.1±0.7	73.8±1.0	65.4±0.9	98.3±0.1	55.1±0.5	57.3±0.3	86.3±0.6
	zipper	92.5±1.0	95.4±0.6	87.9±1.2	97.8±0.2	33.2±1.2	36.5±1.0	93.7±0.3
	Mean	83.6±0.2	82.0±0.3	74.9±0.2	98.2±0.0	36.4±0.3	41.7±0.4	92.5±0.2
4-shot	audiojack	77.7±0.8	72.9±0.8	61.9±0.8	98.6±0.2	31.9±1.3	44.0±0.7	91.3±0.9
	bottle_cap	81.2±0.2	76.4±0.3	76.1±0.5	99.8±0.1	35.0±0.5	39.7±0.2	97.4±0.0
	button_battery	80.5±2.6	86.6±1.7	76.6±1.7	97.8±0.3	52.3±4.2	53.0±3.8	87.5±1.5
	end_cap	76.7±1.2	82.2±1.2	76.3±0.3	98.8±0.1	25.8±1.1	32.2±1.4	95.5±0.1
	eraser	93.5±0.3	92.5±0.1	83.8±0.4	99.8±0.0	40.5±1.1	44.2±0.6	98.7±0.1
	fire_hood	87.3±1.3	78.3±2.1	73.6±1.3	99.6±0.1	32.6±3.9	39.7±2.4	94.1±0.3
	mint	71.5±2.0	72.1±2.4	66.6±1.0	95.8±0.2	21.5±2.3	32.2±1.8	80.6±1.0
	mounts	85.3±0.2	71.1±0.5	76.1±0.4	99.3±0.0	35.1±0.2	40.7±0.4	97.4±0.2
	pcb	84.9±0.5	91.0±0.4	81.1±0.5	99.0±0.1	44.2±0.1	47.7±0.1	94.8±0.1
	phone_battery	89.1±0.4	88.2±0.4	78.0±0.4	97.0±0.1	38.7±0.4	43.6±0.2	95.8±0.2
	plastic_nut	85.9±0.2	79.1±0.5	69.9±0.8	99.4±0.0	38.5±0.6	39.2±0.6	96.9±0.1
	plastic_plug	80.5±1.0	74.8±1.7	65.4±0.7	99.1±0.0	28.2±0.8	35.7±0.5	95.1±0.4
	porcelain_doll	90.9±0.9	84.7±2.2	76.9±0.9	99.6±0.1	35.3±2.9	40.1±2.4	97.8±0.1
	regulator	68.9±0.9	51.1±2.6	50.3±0.4	97.9±0.1	13.2±4.0	22.7±5.7	88.9±0.3
	rolled_strip_base	91.5±1.4	95.6±0.8	88.3±0.9	99.6±0.2	33.3±2.2	44.0±1.3	98.6±0.4
	sim_card_set	97.1±0.1	97.6±0.1	92.0±0.1	99.8±0.0	60.4±0.4	57.7±0.8	97.9±0.1
	switch	84.0±0.5	87.7±0.4	76.2±0.3	96.9±0.0	39.4±0.6	46.3±0.4	94.6±0.1
	tape	93.5±0.3	92.2±0.3	83.8±0.2	99.7±0.1	46.4±0.5	49.6±0.2	97.8±0.1
	terminalblock	87.1±0.4	90.0±0.4	80.1±0.2	99.6±0.0	46.5±1.0	50.0±0.5	98.3±0.2
	toothbrush	78.5±0.3	83.3±0.1	73.1±0.6	95.0±0.6	25.4±1.5	34.7±1.9	78.3±2.2
	toy	73.8±7.2	81.5±4.7	76.6±3.0	90.8±0.3	20.0±0.5	28.8±1.6	82.6±0.4
	toy_brick	74.1±1.4	70.2±1.6	63.2±1.0	97.6±0.1	32.1±1.0	37.3±0.5	84.8±1.0
	transistor1	83.4±0.3	87.1±0.4	79.4±0.1	97.1±0.2	33.3±2.0	39.6±2.6	89.0±0.6
	u_block	86.3±0.4	83.2±0.7	75.3±1.1	99.7±0.0	50.3±0.6	53.3±1.2	97.9±0.1
	usb	87.4±0.1	87.4±0.1	79.0±0.4	99.1±0.1	34.6±0.3	41.0±0.4	96.5±0.0
	usb_adaptor	81.5±0.5	73.9±1.0	71.0±0.2	99.4±0.0	26.6±0.7	32.5±0.9	95.4±0.2
	vcpill	90.2±0.1	88.8±0.1	77.8±0.1	99.1±0.0	58.5±1.7	61.0±0.7	90.8±0.2
	wooden_beads	88.0±0.5	87.6±0.5	78.0±0.8	99.4±0.1	45.3±0.5	48.2±0.4	93.4±0.3
	woodstick	85.0±0.4	75.1±1.1	66.3±0.8	98.4±0.1	55.5±0.5	57.1±0.3	87.2±0.5
	zipper	93.7±0.1	96.0±0.1	89.3±0.2	98.1±0.2	35.4±2.4	38.8±3.5	94.5±0.6
	Mean	84.3±0.2	82.6±0.2	75.4±0.1	98.3±0.1	37.2±0.2	42.5±0.3	93.0±0.2

Table A25: Image anomaly classification and pixel anomaly segmentation results on **Real-IAD** with zero-shot and 1-shot **UniADet**[‡].

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
0-shot	audiojack	70.4	62.6	55.6	97.8	34.5	39.9	86.0
	bottle_cap	82.5	79.6	72.5	99.7	30.7	35.2	98.0
	button_battery	76.9	84.3	74.0	98.2	43.5	43.3	92.0
	end_cap	71.9	78.6	74.9	98.0	25.2	31.7	93.6
	eraser	92.2	90.8	81.4	99.9	56.2	55.1	99.1
	fire_hood	86.2	75.6	71.9	99.7	46.6	48.2	95.8
	mint	71.1	70.5	66.5	99.0	37.2	40.5	89.2
	mounts	87.1	71.0	78.1	99.3	50.0	49.7	98.3
	pcb	80.8	88.1	78.8	97.5	32.0	38.2	93.4
	phone_battery	82.8	82.9	71.6	93.5	45.7	50.8	97.1
	plastic_nut	84.8	77.3	67.6	99.6	52.3	51.8	97.4
	plastic_plug	82.6	74.6	69.3	99.6	32.8	36.9	97.7
	porcelain_doll	91.1	85.0	75.9	99.6	44.4	45.7	98.4
	regulator	78.1	66.1	56.5	98.5	44.5	48.4	89.0
	rolled_strip_base	94.4	97.1	91.0	99.7	58.3	56.3	99.0
	sim_card_set	94.1	95.3	88.1	99.9	69.8	64.7	98.7
	switch	68.9	72.9	70.7	82.6	13.6	19.7	84.8
	tape	95.9	94.8	87.3	99.9	70.1	63.8	99.2
	terminalblock	87.0	90.3	79.3	99.1	51.2	52.6	97.3
	toothbrush	82.4	84.8	77.0	96.4	27.1	33.0	77.8
	toy	68.0	76.3	73.5	88.2	13.5	22.0	79.9
	toy_brick	80.0	74.1	69.3	99.1	50.2	50.1	92.1
	transistor1	86.1	89.9	80.0	96.6	45.4	47.0	88.1
	u_block	84.0	74.4	67.0	99.7	55.0	53.2	97.2
	usb	73.9	73.9	67.0	97.1	23.9	30.2	90.6
	usb_adaptor	77.2	72.5	65.4	99.1	32.1	37.7	93.0
	vcpill	82.1	74.7	70.1	99.3	62.6	59.4	94.2
	wooden_beads	83.3	80.7	71.8	99.6	52.5	53.3	96.1
	woodstick	87.1	75.6	69.3	99.5	63.4	60.8	95.4
	zipper	92.6	96.0	88.2	96.5	28.6	31.6	95.8
	Mean	82.5	80.3	73.7	97.7	43.1	45.0	93.5
1-shot	audiojack	84.9±1.0	80.4±1.4	69.5±1.3	99.4±0.1	53.1±0.6	55.4±0.1	95.3±0.3
	bottle_cap	95.8±0.6	94.9±0.9	88.7±1.3	99.9±0.0	42.9±0.6	46.8±0.4	98.7±0.1
	button_battery	84.1±1.8	88.7±1.2	79.2±1.5	98.1±0.2	52.4±1.5	52.8±0.5	90.6±1.6
	end_cap	87.0±0.3	89.1±0.2	82.3±0.2	99.3±0.1	27.5±0.3	35.0±0.2	97.3±0.1
	eraser	94.4±0.5	93.1±0.3	85.1±0.5	99.8±0.0	52.2±0.2	51.9±0.3	98.8±0.1
	fire_hood	86.7±2.0	76.9±1.5	72.9±3.1	99.6±0.1	45.0±1.8	48.2±0.1	94.7±1.6
	mint	77.6±1.6	78.9±1.5	69.8±1.3	98.2±0.1	35.6±0.2	40.4±0.3	84.5±0.4
	mounts	85.6±0.3	70.0±0.8	75.1±0.4	99.5±0.1	46.6±0.3	48.4±0.6	97.9±0.4
	pcb	88.8±0.8	93.5±0.5	83.5±0.7	99.2±0.1	56.0±0.5	57.4±0.5	95.2±0.3
	phone_battery	93.0±0.5	91.9±0.6	83.0±0.7	99.8±0.0	61.9±0.2	65.8±0.6	98.3±0.2
	plastic_nut	88.8±0.9	82.1±1.5	73.1±1.6	99.8±0.0	52.0±0.4	51.4±0.2	99.0±0.0
	plastic_plug	86.5±0.1	82.7±0.4	73.2±1.0	98.5±0.2	32.0±0.4	44.4±0.4	91.8±0.8
	porcelain_doll	85.9±1.9	78.3±3.0	69.9±2.0	98.5±0.2	25.0±2.1	37.0±2.4	93.7±0.9
	regulator	76.9±1.5	65.4±3.2	55.7±1.8	99.2±0.1	40.4±2.9	48.6±1.3	94.4±0.6
	rolled_strip_base	96.7±0.5	98.3±0.3	93.6±0.5	99.8±0.0	55.3±1.1	57.2±0.5	98.8±0.2
	sim_card_set	95.4±0.7	96.6±0.6	89.5±1.1	99.8±0.1	65.5±0.6	62.4±0.4	97.8±0.1
	switch	90.9±1.1	92.7±0.9	83.7±0.8	97.9±0.0	55.6±0.4	59.6±0.2	96.4±0.1
	tape	96.6±0.3	95.6±0.4	88.5±0.8	99.9±0.0	68.4±0.8	63.3±0.4	98.8±0.1
	terminalblock	95.4±0.5	96.7±0.4	90.6±1.0	99.7±0.1	58.4±0.3	57.1±0.4	98.8±0.1
	toothbrush	83.8±1.6	86.6±1.0	77.9±1.5	95.9±1.1	26.9±3.3	33.1±3.5	86.1±4.3
	toy	76.7±3.7	84.2±2.8	76.5±1.0	93.4±0.9	20.1±0.7	30.6±1.1	87.2±1.0
	toy_brick	81.8±0.7	75.2±1.4	71.6±0.4	97.4±0.3	46.2±1.0	48.8±0.8	84.4±2.5
	transistor1	89.9±3.9	92.7±2.7	84.0±3.7	98.9±0.4	46.9±6.5	48.7±3.2	95.1±1.7
	u_block	90.9±0.9	86.7±1.6	76.3±2.4	99.7±0.1	61.0±3.3	60.5±1.6	97.6±0.8
	usb	91.4±0.7	91.2±0.7	82.7±0.7	99.2±0.0	44.8±0.6	46.5±0.8	97.2±0.1
	usb_adaptor	87.9±0.3	85.7±0.4	75.6±0.5	99.6±0.0	35.3±0.3	39.6±0.4	97.0±0.3
	vcpill	91.4±1.4	89.4±1.9	80.0±1.8	99.1±0.1	70.9±2.2	69.0±1.0	91.4±1.2
	wooden_beads	89.9±0.9	89.2±0.8	79.6±1.0	99.6±0.1	62.3±1.2	61.0±0.4	91.8±0.8
	woodstick	88.1±1.7	78.0±1.9	73.4±1.5	99.3±0.3	63.2±2.2	61.0±1.3	92.6±2.6
	zipper	97.2±0.5	98.3±0.3	93.6±1.1	98.9±0.1	47.9±0.6	52.4±0.8	97.5±0.2
	Mean	88.7±0.5	86.8±0.3	79.3±0.4	98.9±0.0	48.4±0.4	51.1±0.3	94.6±0.4

Table A26: Image anomaly classification and pixel anomaly segmentation results on **Real-IAD** with 2-shot and 4-shot **UniADet**[‡].

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
2-shot	audiojack	86.4±1.3	81.9±1.9	71.1±1.8	99.4±0.0	53.0±0.9	55.0±0.1	95.6±0.2
	bottle_cap	96.4±0.5	95.8±0.6	89.6±1.1	99.9±0.0	43.1±0.1	46.7±0.0	98.8±0.0
	button_battery	87.2±0.9	90.5±0.6	82.2±1.1	98.5±0.1	54.1±1.4	51.4±0.8	92.1±1.1
	end_cap	87.8±1.3	89.7±0.7	83.4±1.5	99.4±0.0	27.0±0.1	34.4±0.0	97.6±0.2
	eraser	95.2±0.4	94.1±0.7	86.2±1.1	99.8±0.1	52.2±0.6	51.7±0.0	99.0±0.3
	fire_hood	87.2±0.3	77.8±0.0	72.5±0.6	99.7±0.0	44.8±0.5	48.2±0.1	95.2±0.2
	mint	78.4±0.2	79.8±0.1	70.0±0.4	98.5±0.1	35.1±2.0	39.8±1.1	86.4±1.2
	mounts	85.6±0.1	69.8±0.4	75.4±0.1	99.6±0.1	45.8±0.7	48.0±0.2	98.0±0.4
	pcb	90.4±1.0	94.4±0.6	85.2±0.7	99.4±0.1	58.0±0.1	58.3±0.1	96.0±0.4
	phone_battery	94.3±0.3	93.0±0.4	84.6±0.2	99.8±0.0	62.7±0.0	66.6±0.6	98.5±0.0
	plastic_nut	90.9±1.9	84.8±1.9	76.9±4.5	99.8±0.1	51.8±0.8	50.6±0.2	99.1±0.1
	plastic_plug	87.2±2.1	83.8±2.8	73.4±3.0	98.5±0.1	30.6±0.3	43.0±0.8	91.6±1.3
	porcelain_doll	85.0±1.8	76.2±2.5	68.4±1.6	98.6±0.2	21.9±1.3	33.2±1.6	93.6±0.6
	regulator	80.6±1.8	70.0±1.3	59.8±3.6	99.5±0.0	45.5±2.7	50.0±0.2	95.9±0.2
	rolled_strip_base	97.2±0.8	98.6±0.5	94.4±0.6	99.8±0.0	54.2±0.9	56.1±0.8	99.0±0.1
	sim_card_set	95.8±0.5	96.8±0.4	90.0±0.7	99.8±0.0	65.4±0.8	62.4±0.4	98.0±0.1
	switch	92.6±0.6	94.0±0.3	85.6±0.8	98.2±0.0	57.0±0.1	60.2±0.1	97.0±0.1
	tape	96.9±0.0	95.8±0.1	88.4±0.5	99.9±0.0	68.0±0.7	62.8±0.4	98.8±0.1
	terminalblock	96.1±0.2	97.2±0.1	91.6±1.2	99.7±0.0	58.7±0.0	57.2±0.2	99.0±0.1
	toothbrush	84.4±0.4	86.8±0.4	79.0±0.1	96.9±0.1	28.0±0.5	34.6±0.8	89.6±1.1
	toy	82.2±0.4	88.2±0.5	78.5±0.7	94.2±1.1	25.8±1.5	34.4±0.4	89.4±1.6
	toy_brick	82.7±1.3	76.8±2.5	72.4±0.6	97.8±0.2	47.2±1.6	49.4±1.8	86.6±1.2
	transistor1	92.4±2.0	94.4±1.5	86.4±2.3	99.1±0.1	50.4±2.1	50.4±1.6	96.4±0.5
	u_block	91.6±2.5	87.6±4.2	77.9±5.5	99.8±0.1	60.6±5.2	59.0±2.7	97.8±1.0
	usb	91.8±0.4	91.4±0.5	82.8±0.6	99.4±0.0	44.9±0.1	46.5±0.4	97.8±0.1
	usb_adaptor	88.2±0.1	85.9±0.0	75.8±0.2	99.6±0.0	34.8±0.5	39.1±0.6	97.1±0.0
	vcpill	92.1±0.4	90.5±1.0	80.8±1.1	99.4±0.1	73.0±1.7	69.3±1.1	93.2±0.8
	wooden_beads	91.4±0.6	91.0±0.6	81.6±0.8	99.6±0.1	63.2±0.8	61.2±0.4	93.2±0.4
	woodstick	88.7±0.6	78.3±0.0	72.9±0.2	99.5±0.0	62.7±0.3	60.2±0.2	94.2±0.4
	zipper	98.0±0.1	98.7±0.1	95.1±0.2	99.0±0.1	49.6±0.3	53.6±0.3	97.8±0.1
	Mean	89.8±0.4	87.8±0.3	80.4±0.6	99.0±0.1	49.0±0.6	51.1±0.1	95.4±0.2
4-shot	audiojack	87.4±0.6	83.0±0.8	72.0±0.7	99.4±0.1	52.8±0.5	54.8±0.1	95.8±0.1
	bottle_cap	96.6±0.0	96.0±0.1	89.6±0.1	99.9±0.0	42.6±0.1	46.2±0.4	98.8±0.0
	button_battery	88.0±1.7	90.8±1.1	83.4±2.2	98.6±0.2	53.7±1.4	49.8±0.0	92.2±1.1
	end_cap	88.0±0.8	89.4±0.5	83.6±0.7	99.5±0.0	25.8±0.1	33.6±0.2	97.6±0.1
	eraser	95.2±0.4	94.0±0.6	86.0±0.6	99.8±0.1	51.7±0.6	51.2±0.1	99.0±0.1
	fire_hood	88.2±0.2	78.9±0.1	73.8±0.6	99.7±0.0	45.4±0.1	48.2±0.1	95.7±0.3
	mint	79.7±1.1	80.9±0.9	71.2±1.2	98.8±0.1	36.1±1.0	39.8±0.5	88.2±0.3
	mounts	86.1±0.6	70.2±0.2	76.4±1.8	99.6±0.1	45.6±0.9	48.0±0.0	98.1±0.1
	pcb	91.2±0.2	94.8±0.1	85.9±0.1	99.5±0.0	59.0±0.1	58.6±0.1	96.4±0.1
	phone_battery	94.7±0.1	93.4±0.4	85.2±0.1	99.8±0.0	63.6±0.7	67.4±0.5	98.6±0.0
	plastic_nut	91.5±1.6	85.5±1.8	77.8±4.2	99.9±0.0	51.6±0.1	50.2±0.4	99.1±0.0
	plastic_plug	89.2±1.1	85.8±1.5	75.6±1.5	98.5±0.1	29.9±0.1	40.8±0.1	91.4±1.3
	porcelain_doll	84.4±1.6	73.6±2.4	68.2±1.5	98.5±0.1	17.5±3.1	26.9±4.5	93.5±0.3
	regulator	83.1±0.6	73.4±1.6	63.3±1.3	99.6±0.0	48.6±1.0	51.6±1.9	96.6±0.1
	rolled_strip_base	97.6±0.1	98.8±0.1	94.8±0.4	99.8±0.0	53.7±1.0	55.6±1.1	99.0±0.1
	sim_card_set	96.2±0.3	97.1±0.3	90.4±0.4	99.8±0.0	64.7±0.3	62.1±0.0	98.0±0.1
	switch	93.6±0.8	94.9±0.9	87.5±1.3	98.4±0.1	57.8±0.0	60.4±0.2	97.3±0.0
	tape	97.6±0.6	96.6±0.9	89.6±2.2	99.9±0.0	67.5±1.1	62.0±0.6	99.0±0.1
	terminalblock	96.4±0.2	97.6±0.1	92.8±0.1	99.8±0.0	58.8±0.2	57.2±0.1	99.0±0.1
	toothbrush	84.5±0.0	87.1±0.1	78.6±0.1	97.2±0.2	28.6±0.4	34.8±0.2	89.6±1.1
	toy	83.6±2.7	88.9±1.9	80.0±1.3	94.7±0.7	26.0±0.8	33.9±0.4	90.0±0.1
	toy_brick	83.4±1.2	78.6±1.6	72.5±1.1	98.1±0.1	49.4±0.2	51.6±0.1	87.1±1.8
	transistor1	94.6±0.8	96.0±0.6	89.0±0.8	99.4±0.0	52.6±0.6	51.5±0.1	97.4±0.1
	u_block	93.2±0.2	90.3±0.3	80.9±0.6	99.8±0.0	62.4±0.2	59.4±0.4	98.2±0.3
	usb	92.4±0.8	92.1±0.8	84.1±1.6	99.5±0.0	45.5±0.7	47.2±0.8	98.0±0.1
	usb_adaptor	87.8±0.2	85.2±0.2	75.6±0.6	99.6±0.0	34.0±0.4	38.2±0.6	97.0±0.1
	vcpill	93.4±0.4	92.2±0.2	82.8±0.6	99.4±0.0	73.8±0.9	69.8±0.5	93.8±0.1
	wooden_beads	91.8±0.5	91.6±0.5	82.4±0.8	99.6±0.1	63.4±0.2	61.0±0.1	93.8±0.4
	woodstick	89.6±0.0	80.0±0.6	74.4±0.4	99.6±0.1	62.9±0.8	60.5±0.3	94.6±0.1
	zipper	98.2±0.2	98.8±0.2	95.5±0.4	99.2±0.1	52.3±3.5	56.8±5.0	98.0±0.4
	Mean	90.6±0.1	88.5±0.1	81.4±0.2	99.2±0.1	49.2±0.1	51.0±0.0	95.7±0.1

Table A27: Image anomaly classification and pixel anomaly segmentation results on **Real-IAD** with zero-shot and 1-shot **UniADet***.

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
0-shot	audiojack	68.1	57.4	54.4	97.7	19.5	24.8	83.8
	bottle_cap	74.9	73.3	65.9	99.4	26.2	30.7	96.7
	button_battery	74.0	81.9	73.7	97.9	42.8	43.7	88.1
	end_cap	71.7	78.7	74.3	96.7	22.9	31.7	88.7
	eraser	89.1	87.8	77.4	99.8	55.5	54.5	98.1
	fire_hood	86.5	76.7	73.0	99.6	42.3	45.6	95.5
	mint	71.1	71.8	65.8	97.9	38.8	43.0	84.1
	mounts	83.0	63.3	74.3	98.6	42.5	46.3	96.7
	pcb	73.0	82.1	77.2	97.7	25.8	30.5	91.7
	phone_battery	84.0	84.4	72.8	90.5	47.2	52.1	95.9
	plastic_nut	89.4	82.5	74.0	99.5	52.0	51.2	96.4
	plastic_plug	83.4	75.6	70.4	99.5	33.2	36.2	97.4
	porcelain_doll	92.0	86.0	78.3	99.7	47.1	47.6	98.4
	regulator	74.1	59.9	52.2	97.1	33.5	39.0	80.4
	rolled_strip_base	93.7	96.7	90.1	99.5	58.5	56.3	98.3
	sim_card_set	95.9	96.8	90.7	99.9	73.8	67.9	98.8
	switch	70.7	76.8	69.4	87.3	14.2	20.5	84.0
	tape	91.8	90.5	80.3	99.6	68.7	63.8	97.2
	terminalblock	84.4	88.7	77.9	98.9	49.6	51.9	96.3
	toothbrush	79.9	83.9	74.4	95.6	25.4	31.4	78.4
	toy	65.9	75.3	73.8	86.0	15.1	22.4	75.2
	toy_brick	82.1	76.2	71.6	98.9	47.4	48.8	93.9
	transistor1	82.8	87.7	78.5	97.6	46.4	48.4	89.8
	u_block	81.4	66.3	67.1	99.4	47.0	48.1	94.8
	usb	71.4	73.3	64.1	97.4	24.2	30.1	89.5
	usb_adaptor	76.4	73.0	64.7	99.0	33.5	38.3	91.4
	vcpill	83.0	80.0	70.1	99.1	65.7	61.9	91.8
	wooden_beads	83.8	83.2	72.1	99.4	58.0	57.9	94.3
	woodstick	85.6	74.1	67.5	99.6	65.2	62.1	95.3
	zipper	91.4	95.5	87.2	96.4	27.3	30.8	95.5
	Mean	81.2	79.3	72.8	97.5	41.6	43.9	91.9
1-shot	audiojack	83.0±1.0	79.6±1.0	69.6±0.9	98.9±0.1	51.0±1.0	56.1±0.8	91.1±0.6
	bottle_cap	88.6±0.5	87.2±0.6	77.1±0.9	99.9±0.1	41.3±0.3	43.2±0.5	98.7±0.1
	button_battery	78.8±2.7	84.9±2.1	76.4±1.3	97.6±0.2	48.9±2.5	48.8±2.4	88.2±1.1
	end_cap	82.3±0.4	84.7±0.2	80.2±0.3	99.0±0.1	22.7±0.5	30.9±0.5	96.2±0.2
	eraser	93.3±0.7	92.0±0.8	83.7±0.6	99.8±0.0	52.8±1.4	52.0±0.9	98.8±0.1
	fire_hood	89.7±1.3	81.9±1.9	76.7±1.6	99.6±0.0	45.1±3.6	46.3±1.0	95.8±0.6
	mint	74.9±1.8	76.4±2.1	67.8±0.9	96.9±0.1	30.8±2.4	39.8±2.4	82.7±0.9
	mounts	84.9±0.2	70.6±0.7	75.2±0.1	99.2±0.1	47.2±0.7	47.4±0.5	97.6±0.0
	pcb	85.3±0.5	91.7±0.3	80.7±0.6	99.1±0.0	55.8±0.3	57.1±0.2	93.7±0.3
	phone_battery	92.3±0.6	90.9±1.0	82.2±1.3	99.8±0.0	58.7±2.0	53.7±2.7	98.4±0.1
	plastic_nut	89.9±1.5	83.4±2.1	74.6±3.0	99.8±0.0	51.0±0.8	50.1±0.3	98.6±0.1
	plastic_plug	85.6±0.1	82.1±0.4	72.5±0.4	99.3±0.1	33.9±0.5	39.2±0.6	96.7±0.2
	porcelain_doll	90.4±0.8	84.6±1.5	75.0±1.1	99.3±0.1	40.2±1.4	44.8±0.9	97.1±0.3
	regulator	73.2±1.9	55.2±2.5	53.9±1.7	98.2±0.2	26.9±2.6	33.5±2.2	88.1±0.9
	rolled_strip_base	96.6±0.4	98.2±0.3	93.5±0.3	99.8±0.0	57.4±1.0	58.4±0.7	99.2±0.1
	sim_card_set	96.1±0.7	97.1±0.5	90.6±1.0	99.8±0.0	69.4±0.6	65.1±0.2	98.2±0.1
	switch	86.0±0.6	89.6±0.4	77.5±0.6	97.2±0.0	54.8±0.1	57.8±0.1	95.0±0.1
	tape	96.2±0.2	95.3±0.2	87.9±0.1	99.8±0.0	66.7±0.7	62.0±0.3	98.4±0.1
	terminalblock	92.0±0.4	94.5±0.3	86.6±0.6	99.7±0.0	58.2±0.4	56.2±0.2	98.9±0.2
	toothbrush	83.8±1.1	86.5±0.6	78.3±1.1	96.6±0.6	30.3±2.4	37.1±1.9	86.1±2.6
	toy	68.2±5.3	73.7±5.9	76.0±1.0	88.5±1.3	6.8±0.7	15.3±2.3	81.7±0.4
	toy_brick	81.2±1.6	76.2±2.7	70.5±1.0	97.4±0.2	42.8±2.0	46.7±1.7	87.7±1.2
	transistor1	86.9±4.8	90.2±3.8	81.3±3.3	98.8±0.3	47.7±6.1	48.7±4.7	94.4±1.5
	u_block	88.4±0.9	83.0±1.7	73.0±1.8	99.7±0.0	59.8±3.7	57.7±3.8	97.3±0.5
	usb	88.3±0.6	88.4±0.4	78.8±0.4	98.7±0.1	43.3±0.2	46.8±0.1	95.1±0.1
	usb_adaptor	88.0±1.2	85.8±1.5	76.1±1.8	99.6±0.1	34.7±0.7	38.4±0.5	97.4±0.3
	vcpill	90.5±0.7	89.7±0.8	79.3±1.2	99.0±0.1	72.7±1.1	69.9±0.9	90.5±0.8
	wooden_beads	90.0±0.4	89.8±0.4	79.4±0.5	99.4±0.1	63.6±0.3	61.7±0.2	92.1±0.3
	woodstick	87.4±1.2	78.0±0.8	69.8±0.8	99.0±0.3	64.0±0.9	62.9±0.8	90.8±2.4
	zipper	96.7±0.2	98.1±0.2	92.6±0.4	99.1±0.1	50.1±1.7	53.8±1.9	97.6±0.3
	Mean	86.9±0.6	85.3±0.5	77.9±0.6	98.6±0.1	47.6±0.5	49.4±0.4	94.0±0.3

Table A28: Image anomaly classification and pixel anomaly segmentation results on **Real-IAD** with 2-shot and 4-shot **UniADet***.

Shot	Categories	Anomaly Classification			Anomaly Segmentation			
		I-AUROC	I-AUPR	I-F1$_{max}$	P-AUROC	P-AUPR	P-F1$_{max}$	P-AUPRO
2-shot	audiojack	83.9±0.7	80.4±0.7	70.2±0.8	99.0±0.1	51.6±0.3	56.8±0.2	91.9±0.5
	bottle_cap	88.4±0.4	87.2±0.5	76.6±0.6	99.9±0.0	41.6±0.3	43.5±0.5	98.8±0.1
	button_battery	83.6±0.6	88.4±0.4	78.6±0.3	98.0±0.1	53.4±1.1	52.6±0.5	90.0±0.7
	end_cap	82.4±1.2	84.6±0.8	80.2±0.8	99.1±0.1	22.2±1.0	30.6±0.8	96.4±0.3
	eraser	93.4±0.3	91.9±0.7	83.8±0.8	99.8±0.0	52.2±1.2	51.8±0.6	98.9±0.0
	fire_hood	90.1±0.5	82.3±1.1	77.2±0.8	99.6±0.1	45.8±1.7	46.8±0.5	96.0±0.2
	mint	76.4±1.1	77.8±1.3	68.4±0.7	97.4±0.1	31.4±2.2	40.6±1.4	85.0±0.6
	mounts	85.1±0.2	71.0±0.3	75.5±0.3	99.4±0.1	47.4±0.2	47.4±0.1	97.9±0.0
	pcb	85.7±0.4	91.9±0.3	81.4±0.5	99.3±0.1	57.4±0.3	57.8±0.2	94.9±0.2
	phone_battery	92.8±0.4	91.2±0.9	83.1±0.5	99.8±0.0	59.0±2.2	54.4±3.1	98.6±0.1
	plastic_nut	90.7±1.3	84.6±1.8	75.8±2.8	99.8±0.0	51.1±0.5	50.2±0.2	98.8±0.1
	plastic_plug	86.2±0.7	82.8±0.4	73.0±0.9	99.3±0.1	33.2±0.4	38.0±0.5	96.7±0.4
	porcelain_doll	90.4±0.7	84.1±0.9	75.2±0.9	99.3±0.1	38.0±0.2	43.1±0.2	96.9±0.2
	regulator	77.9±2.1	62.7±4.5	58.8±3.2	98.7±0.1	34.8±3.3	40.5±2.8	90.5±0.2
	rolled_strip_base	97.0±0.4	98.4±0.2	93.8±0.5	99.8±0.0	57.3±0.8	58.4±0.2	99.2±0.1
	sim_card_set	96.2±0.6	97.2±0.4	90.8±0.8	99.8±0.0	69.4±0.5	64.9±0.2	98.2±0.2
	switch	85.7±0.1	89.2±0.1	76.6±0.2	97.7±0.0	55.4±0.1	57.8±0.2	95.9±0.1
	tape	96.5±0.4	95.7±0.4	88.4±0.6	99.8±0.0	67.2±0.1	62.3±0.1	98.4±0.1
	terminalblock	91.9±0.3	94.5±0.1	86.7±0.0	99.8±0.0	58.6±0.1	56.2±0.2	99.0±0.1
	toothbrush	82.7±0.8	85.6±0.4	77.6±1.1	96.6±0.9	29.2±3.7	35.4±2.3	86.4±5.0
	toy	79.6±1.0	86.3±0.7	77.2±0.4	90.0±0.3	25.4±0.8	34.8±0.8	83.9±0.4
	toy_brick	83.6±0.4	79.8±0.1	72.4±1.1	97.8±0.1	46.4±0.2	49.6±0.2	88.8±0.1
	transistor1	88.2±3.0	91.3±2.4	82.0±2.2	98.9±0.2	49.9±2.7	50.9±1.3	95.1±0.8
	u_block	88.8±1.6	83.8±3.0	74.2±2.8	99.8±0.0	61.6±4.2	59.4±4.2	97.9±0.1
	usb	88.8±0.8	88.9±0.8	79.4±0.9	99.0±0.1	43.8±0.4	47.2±0.4	96.0±0.3
	usb_adaptor	88.3±0.8	86.5±0.8	77.0±1.1	99.7±0.0	34.8±0.5	38.4±0.2	97.6±0.2
	vcpill	90.8±0.5	90.2±0.6	80.2±0.9	99.2±0.1	73.4±0.2	70.6±0.1	91.3±0.3
	wooden_beads	90.7±0.7	90.4±0.8	80.4±1.6	99.4±0.1	64.3±0.4	61.8±0.4	93.0±0.0
	woodstick	88.4±0.6	79.2±0.2	71.1±0.8	99.2±0.1	64.7±0.3	62.8±0.8	92.4±0.5
	zipper	97.2±0.6	98.4±0.4	93.2±1.2	99.2±0.1	50.8±1.5	54.4±1.3	97.8±0.1
	Mean	87.2±1.2	85.0±2.2	77.4±1.9	98.9±0.2	47.0±2.9	49.0±2.3	94.8±0.2
4-shot	audiojack	84.0±0.7	80.8±0.5	70.6±0.5	99.1±0.2	51.8±0.6	57.1±0.3	92.6±1.0
	bottle_cap	87.8±0.3	86.7±0.3	75.5±0.5	99.9±0.0	41.6±0.2	43.6±0.3	98.8±0.0
	button_battery	85.1±1.6	89.2±1.0	80.0±1.7	98.3±0.3	55.3±2.9	53.5±1.5	91.0±1.4
	end_cap	83.6±0.2	85.3±0.3	80.9±0.4	99.2±0.1	22.6±0.6	31.0±0.5	96.8±0.2
	eraser	93.4±0.5	91.9±0.6	83.7±0.5	99.8±0.1	52.1±1.6	51.8±0.8	98.9±0.1
	fire_hood	90.7±0.8	82.9±1.5	77.8±1.2	99.7±0.0	46.3±1.8	47.1±0.8	96.4±0.1
	mint	78.1±1.6	79.3±1.7	69.8±1.1	97.9±0.2	33.8±3.2	41.4±1.3	86.7±0.9
	mounts	85.2±0.3	70.1±0.7	76.8±0.6	99.5±0.1	46.9±0.1	47.2±0.2	98.1±0.1
	pcb	85.6±0.3	91.9±0.1	81.3±0.3	99.5±0.0	58.0±0.1	57.8±0.2	95.6±0.1
	phone_battery	93.0±0.4	91.9±0.4	83.5±0.4	99.8±0.0	61.3±0.4	57.4±0.7	98.7±0.1
	plastic_nut	92.1±1.2	86.7±1.8	78.4±2.9	99.8±0.0	51.3±0.2	50.3±0.3	98.9±0.1
	plastic_plug	87.1±0.7	83.6±0.7	73.9±1.2	99.4±0.0	32.8±0.6	37.3±0.9	96.9±0.2
	porcelain_doll	90.4±0.7	83.5±1.6	75.4±0.9	99.2±0.1	37.3±2.6	42.8±2.8	96.7±0.1
	regulator	79.8±0.5	67.5±4.3	60.7±1.0	98.9±0.1	39.9±1.3	45.1±1.2	91.7±0.3
	rolled_strip_base	97.3±0.2	98.6±0.1	94.0±0.5	99.8±0.1	57.4±0.3	58.6±0.3	99.4±0.1
	sim_card_set	96.4±0.3	97.3±0.2	90.9±0.5	99.8±0.0	69.3±0.9	65.1±0.5	98.3±0.2
	switch	85.2±0.2	88.9±0.2	76.5±0.1	98.0±0.1	55.8±0.1	58.0±0.1	96.3±0.2
	tape	96.8±0.5	95.9±0.6	89.0±1.4	99.9±0.1	67.1±0.3	62.1±0.1	98.4±0.1
	terminalblock	91.9±0.3	94.6±0.2	86.9±0.5	99.8±0.0	59.0±0.2	56.5±0.1	99.1±0.1
	toothbrush	83.6±0.5	86.5±0.4	77.9±0.4	97.2±0.3	31.5±1.9	37.7±1.1	88.3±1.7
	toy	81.7±3.4	87.7±2.4	78.7±2.2	89.6±0.8	25.3±0.4	34.6±0.5	84.3±0.1
	toy_brick	82.9±1.0	79.0±1.3	71.9±1.0	97.9±0.2	48.3±1.5	50.8±1.1	89.1±1.0
	transistor1	90.8±0.6	93.4±0.4	83.9±0.6	99.3±0.1	53.0±1.0	52.4±0.5	96.7±0.3
	u_block	88.7±1.4	83.9±2.8	74.2±2.5	99.8±0.0	61.9±4.3	59.9±4.2	97.9±0.2
	usb	88.4±0.7	88.8±0.6	79.3±0.7	99.2±0.1	44.1±0.4	47.4±0.5	96.7±0.3
	usb_adaptor	86.8±0.8	84.6±0.8	75.0±1.3	99.7±0.0	33.5±0.2	37.5±0.4	97.5±0.1
	vcpill	91.2±0.1	90.7±0.2	80.9±0.1	99.2±0.1	74.4±0.5	71.1±0.3	92.0±0.3
	wooden_beads	91.1±0.5	90.8±0.5	81.0±0.8	99.4±0.1	64.4±0.4	61.8±0.3	93.0±0.4
	woodstick	89.1±0.1	80.4±0.4	72.2±0.2	99.3±0.1	65.3±0.5	63.0±0.4	92.9±0.4
	zipper	97.7±0.2	98.7±0.1	94.4±0.5	99.3±0.1	53.2±1.8	57.5±3.0	98.0±0.3
	Mean	88.5±0.2	87.0±0.2	79.2±0.3	98.9±0.1	49.8±0.2	51.2±0.2	95.2±0.1