AdaptCLIP: Adapting CLIP for Universal Visual Anomaly Detection

Bin-Bin Gao*†[1], Yue Zhou*[2,3,4], Jiangtao Yan[1], Yuezhi Cai[2], Weixi Zhang[2], Meng Wang[2], Jun Liu[1], Yong Liu[1], Lei Wang[2], Chengjie Wang†[1,5]

[1]Tencent Youtu Lab, China
[2]Siemens AG, China
[3]Technical University of Munich, Germany
[4]Munich Center for Machine Learning, Germany
[5]Shanghai Jiao Tong University, China
csgaobb@gmail.com, yue.zhou@tum.de
{yuezhi.cai, weixi.zhang, mengwang.wm, wangleiwl}@siemens.com
{jiangtyan, juliusliu, choasliu, jasoncjwang}@tencent.com

Abstract

Universal visual anomaly detection aims to identify anomalies from novel or unseen vision domains without additional fine-tuning, which is critical in open scenarios. Recent studies have demonstrated that pre-trained vision-language models like CLIP exhibit strong generalization with just zero or a few normal images. However, existing methods struggle to design prompt templates, handle complex token interactions, or require fine-tuning on target domains, resulting in limited flexibility. In this work, we present a simple yet effective AdaptCLIP based on two key insights. First, adaptive visual and textual representations should be learned alternately rather than jointly. Second, comparative learning between query and normal image prompt should incorporate both contextual and aligned residual features, rather than relying solely on residual features. AdaptCLIP treats CLIP models as a foundational service, adding only three simple adapters, visual adapter, textual adapter, and prompt-query adapter, at its input or output ends. AdaptCLIP supports zero-/few-shot generalization across domains and provides a training-free approach on target domains once trained on a base dataset. AdaptCLIP achieves state-of-the-art performance on 12 anomaly detection benchmarks from industrial and medical domains, significantly outperforming existing competitive methods.

Code — https://github.com/gaobb/AdaptCLIP

Introduction

Universal visual anomaly detection (AD) detects anomaly images and segments anomaly pixels from novel or unseen visual objects after learning a single model on a base or seen dataset. This is a more challenging task as it requires strong generalization when facing cross-domain scenarios. Meanwhile, it is a more practical topic as people are more interested in fast adaptability, especially in low data regimes, i.e., few-shot and even zero-shot. For example, in medical image diagnosis and industrial visual quality inspection, it is difficult to collect a large-scale dataset due to inherent scarcity and privacy protection. Recently, developing universal visual AD has attracted increasing attention because existing unsupervised ADs with either separated (Roth et al. 2022; Deng and Li 2022; Liu et al. 2023b) or unified models (You et al. 2022; Gao 2024a) perform poorly in unseen objects despite promising performance on seen objects.

To address this fragmentation, recent works have attempted to design universal models to recognize anomalies for unseen objects. They typically build on vision-language models, i.e., CLIP (Radford et al. 2021). WinCLIP (Jeong et al. 2023) computes anomaly scores on dense patch windows and brings large computational costs and memory burden. AnomalyCLIP (Zhou et al. 2024) learns class-agnostic prompt embeddings to align patch-wise tokens, thus avoiding dense window operations. It further refines vanilla CLIP by concatenating learnable tokens to intermediate layers of

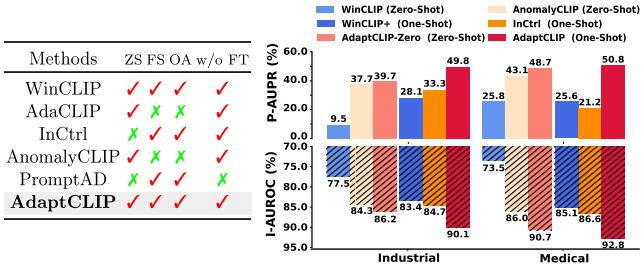

Figure 1: Comparison of state-of-the-art and **AdaptCLIP**. **AdaptCLIP** supports zero-/few-shot (ZS and FS) visual AD across domains without any dataset-specific fine-tuning (FT). It only adds simple adapters at CLIP's input or output ends, thus preserving CLIP's original ability (OA). The one-shot **AdaptCLIP** achieves the best performance on 12 AD benchmarks covering industrial and medical domains. Moreover, the zero-shot **AdaptCLIP** is also significantly better than existing zero-shot and even some one-shot approaches. The detailed results are in Tables 1 and 2.

*These authors contributed equally.
†Corresponding authors.

CLIP. AdaCLIP (Cao et al. 2024) further integrates visual knowledge into textual prompt embeddings. However, they destroy inherent representations of CLIP since learnable tokens are inserted into the middle layers of CLIP. *We want to explore whether we can achieve the same or even better AD performance while maintaining the original ability.*

In contrast, humans perceive anomalies when an input significantly deviates from those normal patterns stored in our brains. There is evidence to support this point in neuroscience (Rao and Ballard 1999). PatchCore (Roth et al. 2022) builds a memory bank storing normal features and PaDiM (Defard et al. 2021) learns a multivariate Gaussian distribution of normal features. At inference, anomalies are recognized by comparing input features with the memory bank or the learned distribution. However, these methods usually require a certain number of normal images and thus are limited in universal (i.e., open-world) scenarios. Recent works, e.g., InCtrl (Zhu and Pang 2024), PromptAD (Li et al. 2024b) and MetaUAS (Gao 2024b), have studied how to further improve performance with few-shot normal images. However, InCtrl only considers anomaly classification, PromptAD needs to learn a new model for each class, and MetaUAS only focuses few-shot scenarios. Different from them, *we want to comprehensively explore a universal AD model, aiming to detect any anomalies in image-/pixel-level from cross-domains without any training on target domains.*

Toward this end, we propose a simple but effective universal visual AD framework, called AdaptCLIP. *The philosophy of AdaptCLIP is that "less and simpler could be better", and it contains two key insights designed by three adapters:* First, adaptive visual and textual representations should be learned alternately rather than jointly. Second, comparative learning between query and the corresponding normal image prompt should incorporate both contextual and aligned residual features, rather than relying solely on residual features. Our contributions are summarized as follows.

- We propose a *simple but effective universal visual anomaly detection framework* based on visual-language CLIP models, which is capable of detecting any visual anomalies at image- and pixel-level from cross-domain datasets with a training-free manner on target domains.
- We propose visual and textual adapters, and find that *they should alternately adapt visual and textual representation* guided by the powerful vision-language representations from CLIP models.
- We propose a prompt-query adapter that aims to *capture meta-perceptual capabilities based on the joint distribution of query contextual features and the aligned residual features* between query image and the corresponding normal image prompt.
- AdaptCLIP outperforms zero- and few-shot AD methods on 8 industrial and 4 medical benchmarks, as shown in Fig. 1. Meanwhile, AdaptCLIP possesses simpler adapters, fewer parameters, and competitive efficiency.

Related Works

Unsupervised ADs target to identify anomalies given sufficient normal training images. Most unsupervised AD methods can be roughly grouped into three categories: embedding-, discrimination-, and reconstruction-based methods. Embedding-based methods, such as PaDiM (Defard et al. 2021), MDND (Rippel, Mertens, and Merhof 2021), PatchCore (Roth et al. 2022), CS-Flow (Rudolph et al. 2022), and PyramidFlow (Lei et al. 2023), assume that offline features extracted from a pre-trained model preserve discriminative information and thus help to separate anomalies from normal samples. Discrimination-based methods, such as CutPaste (Li et al. 2021), DRAEM (Zavrtanik, Kristan, and Skočaj 2021a), SimpleNet (Liu et al. 2023b) and AnoGen (Gui et al. 2024), typically convert unsupervised AD to supervised ones by introducing pseudo (synthesized) anomaly samples. Reconstruction-based ADs, such as autoencoder (Vincent et al. 2008; Bengio et al. 2013; Gong et al. 2019; Hou et al. 2021), generative adversarial networks (Perera, Nallapati, and Xiang 2019; Yan et al. 2021; Zaheer et al. 2020), and reconstruction networks (Zavrtanik, Kristan, and Skočaj 2021b; Ristea et al. 2022; Liu et al. 2023a), assume that anomalous regions should not be able to be properly reconstructed and thus result in high reconstruction errors since they do not exist in normal training samples. The recent knowledge distillation (Bergmann et al. 2020; Wang et al. 2021b,a; Salehi et al. 2021; Deng and Li 2022) or feature reconstruction methods (You et al. 2022; Zhao 2023; Yao et al. 2023a; Gao 2024a) train a student or reconstruction network to match a fixed pre-trained teacher network and achieve a good balance between effectiveness and efficiency. However, these methods are limited to recognizing anomalies of seen classes but often perform poorly on unseen classes. For a novel scenario, people have to collect sufficient normal images first and then retrain a model. This is inefficient and lacks rapid adaptability.

Zero-Shot ADs have achieved impressive performance by utilizing large vision-language models, e.g., CLIP (Radford et al. 2021). WinCLIP (Jeong et al. 2023) designs two-class textual prompts and introduces multi-scale patch windows for accurate anomaly segmentation. It brings large computational costs and memory burden, limiting high-resolution input or large pre-trained models. AnomalyCLIP (Zhou et al. 2024) learns class-agnostic prompt embeddings to align patch-wise tokens thus avoiding dense window operation. In addition, AnomalyCLIP refines vanilla CLIP representation by appending some learnable tokens to the middle layer of CLIP. Recently, AdaCLIP (Cao et al. 2024) and VCP-CLIP (Qu et al. 2024) utilize similar ideas and further integrate visual knowledge into textual prompt embeddings. We argue that these additional operations make models more complex and may hurt the original capabilities of CLIP. Instead of visual-language models, ACR (Li et al. 2023) and MuSc (Li et al. 2024a) perform zero-shot AD only requiring batch-level and full testing images, but they may be limited in privacy protection scenarios. Different from these methods, we explore whether the same or even better AD performance is achieved while retaining the original ability of CLIP without any information on test data distribution.

Few-Shot ADs mainly pay attention to learning or using only a limited number of normal images, such as

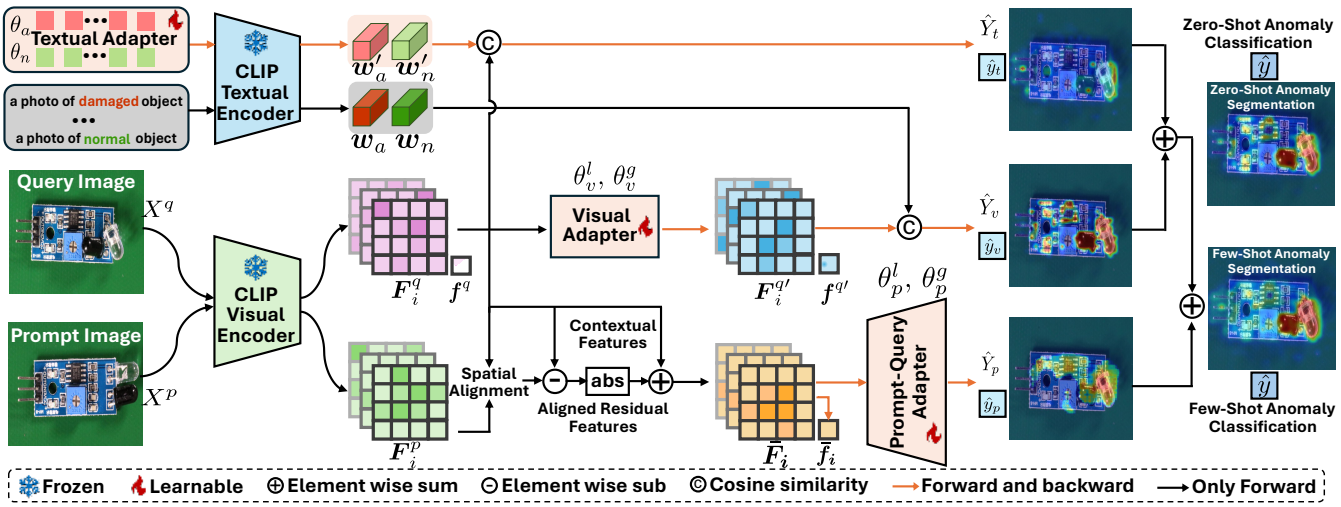

Figure 2: **AdaptCLIP** consists of three pluggable adapters, i.e., visual adapter, textual adapter, and prompt-query adapter. First, the first two adapters alternately learn visual and textual representations for zero-shot anomaly detection. The prompt-query adapter further learns a comparison ability between query image and its corresponding normal prompt for few-shot anomaly detection. Once trained, it can segment any visual anomalies providing only few-shot and even zero-shot normal image prompts.

TDG (Sheynin, Benaim, and Wolf 2021), RegAD (Huang et al. 2022), GraphCore (Xie et al. 2023) and FastRecon (Fang et al. 2023). Some works (Ding, Pang, and Shen 2022; Yao et al. 2023b) consider another few-shot setting where a limited number of samples is given from anomaly images. The performance of these methods lags behind unsupervised ADs. Recently, few-shot AD performance has been improved significantly by visual-language models. WinCLIP+ (Jeong et al. 2023) is the first work to apply CLIP models to few-shot AD, which stores normal tokens into a memory bank, then computes anomaly maps compare between query tokens and the memory. APRIL-GAN (Chen, Han, and Zhang 2023) uses an auxiliary anomaly dataset to fine-tune the CLIP model for better zero-/few-shot performance. InCtrl (Zhu and Pang 2024) further integrates multi-level information to learn a holistic scoring function for anomaly classification. It does not consider pixel-level anomaly segmentation. PromptAD (Li et al. 2024b) introduces the concept of explicit anomaly margin, which mitigates the training challenge caused by the absence of anomaly training images. However, it requires re-training models when applied to target datasets. Recent MeatUAS (Gao 2024b) demonstrated that a pure vision model can achieve strong one-shot performance. Unfortunately, it cannot be extended to zero-shot scenarios. In contrast, we explore jointly optimizing anomaly classification and segmentation in a unified model, which can quickly adapt to novel scenarios only given zero/few-shot normal image prompts, not involving additional re-training.

Methods

Problem Formulation: Our objective is to learn a universal AD model that detects any anomalies from diverse domains without any training on target datasets. Thus, a reasonable assumption is that there is a different distribution between training and testing sets. Formally, let $\mathcal{D}_{\text{base}} = \{X_i, Y_i, y_i\}_{i=1}^{N}$ be a training dataset, that consists of N normal and anomalous images, $X_i \in \mathcal{R}^{h\times w\times 3}$ is the i-th image, and $Y_i \in \mathcal{R}^{h\times w}$ and $y_i = \{0, 1\}$ is the corresponding anomaly mask and anomaly label, with $y_i = 0$ indicates normal and $y_i = 1$ signifies anomaly. The testing set $\mathcal{T}$ may consist of multiple different domains with various objects and anomaly types. Here, we denote the t-th novel domain as $\mathcal{D}^t_{\text{novel}} = \{X_i, Y_i, y_i\}_{i=1}^{N_t}$. Under a few-shot setting, a few normal images $\mathcal{P}_c = \{X_i\}_{i=1}^{k}$ are randomly drawn from each class of the target domain, where c is the class index and k is typically a small number, e.g., $k = \{1, 2, 4\}$. It is worth noting that $\mathcal{P}_c$ is only available during inference, and cannot be used in any way during training phase.

Overview: As illustrated in Fig. 2, image anomaly score and pixel anomaly map can be obtained using textual and visual adapters for zero-shot scenarios. For few-shot scenarios, anomaly score and map are derived by integrating predictions from zero-shot and prompt-query adapters.

Revisiting CLIP for Anomaly Detection

For a query image $X^q \in \mathcal{R}^{h\times w\times 3}$, we feed it to visual encoder $\mathcal{F}(\cdot)$ and obtain local patch tokens $\{\boldsymbol{F}^q_i \in \mathcal{R}^d\}_{i=1}^{hw/p^2}$ and global image token $\boldsymbol{f}^q \in \mathcal{R}^d$, where p is patch size. WinCLIP (Jeong et al. 2023) introduces two-class prompts describing normal and abnormal states. For example, "a photo of a normal object" and "a photo of a damaged object". In practical application, one could design multiple textual descriptions for normal and abnormal states. Feeding these normal and abnormal descriptions to textual encoder $\mathcal{T}(\cdot)$, we can obtain the embeddings of normal $\boldsymbol{w}_n \in \mathcal{R}^d$ and abnormal $\boldsymbol{w}_a \in \mathcal{R}^d$. The pixel-level anomaly map is computed by measuring the cosine similarities between all

patch tokens and the textual embeddings:

$$\hat{Y} = \big[\frac{\exp(\langle \boldsymbol{w}_a, \boldsymbol{F}_i^q \rangle)}{\exp(\langle \boldsymbol{w}_a, \boldsymbol{F}_i^q \rangle) + \exp(\langle \boldsymbol{w}_n, \boldsymbol{F}_i^q \rangle)}\big], \quad (1)$$

where $\langle \cdot \rangle$ represents the cosine similarity, and $[\cdot]$ means that all patch-wise prediction scores are rearranged according to their spatial positions and interpolated to the original input resolution. Replacing $\boldsymbol{F}_i^q$ with $\boldsymbol{f}^q$ in Eq. 1, we can obtain an image-level anomaly score $\hat{y}$ for X^q,

$$\hat{y} = \frac{\exp(\langle \boldsymbol{w}_a, \boldsymbol{f}^q \rangle)}{\exp(\langle \boldsymbol{w}_a, \boldsymbol{f}^q \rangle) + \exp(\langle \boldsymbol{w}_n, \boldsymbol{f}^q \rangle)}. \quad (2)$$

AdaptCLIP with Alternating Learning

To adapt CLIP for universal visual anomaly detection, we design visual and textual adapters to alternately learn visual and textual representations. Specifically, the visual adapter learns adaptive visual tokens ($\boldsymbol{F}_i^{q\prime}$ and $\boldsymbol{f}^{q\prime}$) when fixing two-class static textual embeddings ($\boldsymbol{w}_a$ and $\boldsymbol{w}_n$), while the textual adapter learns two-class textual prompt embeddings ($\boldsymbol{w}'_a$ and $\boldsymbol{w}'_n$) when fixing visual tokens ($\boldsymbol{F}_i^q$ and $\boldsymbol{f}^q$).

Visual Adapter adapts vision tokens ($\boldsymbol{F}_i^q$ and $\boldsymbol{f}^q$) with fixed textual embeddings ($\boldsymbol{w}_a$ and $\boldsymbol{w}_n$). It consists of two branches, global and local, which transform global image tokens and local patch tokens, respectively. Architecturally, the global and local branches are implemented using a simple residual multi-layer perception (MLP), i.e.,

$$\boldsymbol{F}_i^{q\prime} = \boldsymbol{F}_i^q + \text{MLP}(\boldsymbol{F}_i^q; \theta_v^l); \boldsymbol{f}^{q\prime} = \boldsymbol{f}^q + \text{MLP}(\boldsymbol{f}^q; \theta_v^g), \quad (3)$$

where θ_v^l and θ_v^g are learnable parameters. Replacing $\boldsymbol{F}_i^q$ and $\boldsymbol{f}^q$ in Eqs. 1 and 2 with $\boldsymbol{F}_i^{q\prime}$ and $\boldsymbol{f}^{q\prime}$, we obtain pixel-level anomaly map $\hat{Y}_v$ and image-level anomaly score $\hat{y}_v$.

Textual Adapter aims to directly learn two-class prompts $\boldsymbol{\theta}_a, \boldsymbol{\theta}_n \in \mathcal{R}^{r \times d}$ without prompt templates, where $r > 0$ is the length of prompts. We feed them into the frozen textual encoder $\mathcal{T}(\cdot)$ of CLIP, and obtain the corresponding embeddings $\boldsymbol{w}'_a$ and $\boldsymbol{w}'_n$, i.e.,

$$\boldsymbol{w}'_a = \mathcal{T}(\boldsymbol{\theta}_a), \boldsymbol{w}'_n = \mathcal{T}(\boldsymbol{\theta}_n). \quad (4)$$

Then, we replace the static $\boldsymbol{w}_a$ and $\boldsymbol{w}_n$ in Eqs. 1 and 2 with the learnable prompt embeddings $\boldsymbol{w}'_a$ and $\boldsymbol{w}'_n$ to derive local and global anomaly predictions, $\hat{Y}_t$ and $\hat{y}_t$.

Alternating Learning or Joint Learning? A possible question is whether we can learn visual and textual representations jointly. That is, in Eqs. 1 and 2, we simultaneously replace fixed textual embeddings and visual tokens with learnable prompt embeddings ($\boldsymbol{w}'_a$ and $\boldsymbol{w}'_n$) and adaptive visual tokens ($\boldsymbol{F}_i^{q\prime}$ and $\boldsymbol{f}^{q\prime}$). Indeed, this joint alignment mechanism is successful when a large-scale image-text dataset is available. However, we empirically find that it does not work well in the AD field, as shown in Tab. 5 (Lines 3 vs. 4). This is not surprising because the available training data scale is still relatively small and lacks fine-grained textual annotations. The joint learning easily overfits and leads to poor generalization on novel datasets. In contrast, the alternating learning helps us fully utilize the prior knowledge of CLIP and thus improve cross-domain generalization.

AdaptCLIP with Comparative Learning

Compared to static or learnable textual prompts, using a normal image as a visual prompt is more intuitive. Therefore, we expect to learn a comparison ability between a query image X^q and its corresponding normal prompt X^p, which generalizes well to unseen objects. We find that applying multi-layer features yields better results. For simplicity, we use a single-layer feature in the following.

Spatial Alignment: A simple way is to directly measure their difference by the absolute value of their residual feature, that is $|\boldsymbol{F}_i^q - \boldsymbol{F}_i^p|$, where $\boldsymbol{F}_i^q$ and $\boldsymbol{F}_i^p$ are the patch token of X^q and X^p, respectively. It may fail if the query and prompt images are not aligned in pixel space (e.g., due to rotation and translation). Therefore, we have to align query and prompt tokens for effective comparison. For any query token $\boldsymbol{F}_i^q$, we search the nearest one among all normal tokens $\{\boldsymbol{F}_j^p\}_{j=1}^{hw/p^2}$ using euclidean distance:

$$\boldsymbol{F}_i^{p\prime} = \boldsymbol{F}_k^p, k = \arg\min_j \|\boldsymbol{F}_i^q - \boldsymbol{F}_j^p\|_2. \quad (5)$$

Then, we take $\boldsymbol{F}_i^{p\prime}$ as aligned prompt token of $\boldsymbol{F}_i^q$. Now, we can derive the aligned residual feature, i.e., $|\boldsymbol{F}_i^q - \boldsymbol{F}_i^{p\prime}|$.

Joint contextual and aligned residual feature: The aligned residual feature effectively highlights differences or anomaly regions. However, it may lose contextual information or introduce noise. Intuitively, contextual information is critical for identifying anomalies. Therefore, we aggregate the original query tokens and the aligned residual features by an element-wise sum,

$$\bar{\boldsymbol{F}}_i = \boldsymbol{F}_i^q + |\boldsymbol{F}_i^q - \boldsymbol{F}_i^{p\prime}|. \quad (6)$$

Prompt-Query Adapter: The ultimate goal is to achieve pixel-level anomaly segmentation and image-level anomaly classification. Therefore, we propose a lightweight segmentation head $\mathcal{G}(\cdot; \theta_p^l)$ to learn anomaly segmentation based on the joint feature $\bar{F}$:

$$\hat{Y}_p = \mathcal{G}(\bar{F}; \theta_p^l), \quad (7)$$

where θ_p^l is its parameters. Specifically, the segmentation head consists of several transposed convolution blocks following a 1×1 convolution layer. Here, each transposed convolution block upsamples input feature by $2\times$, and it is composed of a 3×3 convolution, a BatchNorm, a ReLU, and a 2×2 deconvolution.

Meanwhile, we need to obtain a global image-level prediction. First, we perform average-pooling and max-pooling on the joint feature $\bar{F}$ along the spatial dimension and then take their weighted average as the global image representation. Then, a simple MLP is used to map the global feature to an image-level prediction score:

$$\hat{y}_p = \text{MLP}\big((\text{AvgPool}(\bar{F}) + \text{MaxPool}(\bar{F}))/2; \theta_p^g\big), \quad (8)$$

where θ_p^g is the parameter.

Training and Inference

During training, we use cross-entropy loss for global image anomaly classification, and Focal and Dice losses for

Shots	Methods	Industrial									Medical		
		MVTec	VisA	BTAD	MVTec3D	DTD	KSDD	MPDD	Real-IAD	AVG	Br35H	Covid	AVG
0	WinCLIP (Jeong et al. 2023)	90.4	75.5	68.2	69.4	95.1	92.9	61.5	67.0	77.5	80.5	66.4	73.5
	AdaCLIP† (Cao et al. 2024)	90.7	81.7	89.9	76.2	92.7	96.6	64.0	73.3	83.1	96.7	69.4	83.0
	AnomalyCLIP (Zhou et al. 2024)	91.6	82.0	88.3	73.9	93.9	97.8	77.5	69.5	84.3	94.2	77.7	86.0
	AdaptCLIP-Zero	93.5	84.8	91.0	78.6	96.0	98.1	73.6	74.2	86.2	94.8	86.5	90.7
1	WinCLIP+ (Jeong et al. 2023)	93.6±0.4	80.0±2.4	84.4±1.5	74.1±0.4	97.9±0.2	93.8±0.4	69.3±2.9	74.7±0.2	83.4	80.1±2.1	90.1±3.6	85.1
	InCtrl (Zhu and Pang 2024)	91.3±0.4	83.2±2.4	88.5±0.4	75.3±1.3	97.9±0.3	92.0±0.9	73.0±2.7	76.6±0.0	84.7	83.9±6.4	89.2±5.3	86.6
	AnomalyCLIP+ (Zhou et al. 2024)	95.2±0.2	86.1±0.7	88.5±0.8	76.7±2.1	98.0±0.2	97.5±0.3	83.4±2.6	78.2±0.0	88.0	90.8±5.1	87.3±2.6	89.1
	AdaptCLIP	94.5±0.5	90.5±1.2	93.4±0.0	81.7±1.5	98.0±0.0	96.9±0.3	83.8±2.2	81.8±0.3	90.1	93.7±2.4	91.8±2.5	92.8
2	WinCLIP+ (Jeong et al. 2023)	94.5±1.0	82.7±1.0	85.8±1.8	74.3±0.3	98.1±0.2	93.8±0.2	69.3±2.3	76.1±0.1	84.3	81.6±0.6	91.8±2.5	86.7
	InCtrl (Zhu and Pang 2024)	91.8±0.9	86.3±1.4	86.2±2.0	75.4±0.5	98.3±0.2	91.6±0.9	74.2±1.8	78.5±0.0	85.3	86.1±1.7	89.7±5.1	87.9
	AnomalyCLIP+ (Zhou et al. 2024)	95.4±0.1	87.8±0.5	89.2±1.1	78.3±1.3	98.2±0.1	97.9±0.2	83.4±1.5	78.3±0.0	88.6	91.5±4.0	89.3±2.7	90.4
	AdaptCLIP	95.7±0.6	92.2±0.8	93.4±0.2	82.9±1.1	98.3±0.0	97.2±0.0	84.4±0.7	82.9±0.2	90.8	94.0±1.7	94.9±0.9	94.5
4	WinCLIP+ (Jeong et al. 2023)	95.3±0.1	84.3±0.6	87.8±0.8	75.7±0.3	98.2±0.0	94.0±0.2	71.2±1.6	77.0±0.0	85.4	82.3±0.4	92.9±2.1	87.6
	InCtrl (Zhu and Pang 2024)	93.1±0.7	87.8±0.2	67.5±2.4	78.1±1.1	97.7±0.1	91.6±0.9	78.6±2.3	81.8±0.0	84.5	89.1±1.2	91.4±4.1	90.3
	AnomalyCLIP+ (Zhou et al. 2024)	96.1±0.1	88.8±0.5	90.5±1.2	79.2±1.3	98.4±0.1	97.8±0.1	86.3±1.8	78.4±0.0	89.4	91.1±4.4	91.4±3.0	91.3
	AdaptCLIP	96.6±0.3	93.1±0.2	93.3±0.3	84.2±0.6	98.5±0.1	97.0±0.2	86.8±1.1	83.9±0.2	91.7	93.7±2.0	95.8±0.9	94.8

Table 1: Image-level anomaly classification comparisons with AUROC metric on industrial and medical domains. The best and second-best results are highlighted in red and blue, respectively. The superscript† indicates that the results are our re-implementation with the same training and testing protocol as AnomalyCLIP and our AdaptCLIP. Note that the results are averaged over all categories on each dataset and the full results of each category are presented in `Appendix`, the same below.

local patch anomaly segmentation, which is the same as AnomalyCLIP (Zhou et al. 2024). For zero-shot inference, we average the predictions from visual and textual adapters. For few-shot inference, we fuse (i.e., average) all results from three adapters, i.e., prompt-query, visual and textual adapters, as the final predictions of AdaptCLIP. The PyTorch pseudocode is shown for the pixel-level inference of AdaptCLIP in `Appendix`. Note that we omit the image-level inference since it can be easily obtained by replacing local patch tokens with global image token.

Experiments

Experimental Setup

Datasets: We comprehensively evaluate AdaptCLIP on multiple datasets from industrial and medical domains. For industrial domain, we use MVTec (Bergmann et al. 2019), VisA (Zou et al. 2022), BTAD (Mishra et al. 2021), MVTec3D (Bergmann. et al. 2022), DTD (Aota, Tong, and Okatani 2023), KSDD (Tabernik et al. 2020), MPDD (Jezek et al. 2021), and large-scale Real-IAD (Wang et al. 2024). In medical domain, we utilize brain tumor detection datasets, Br35H (Hamada 2020) and COVID-19 (Rahman et al. 2021), as well as gastrointestinal polyp datasets, Kvasir (Jha et al. 2020) and Endo (Vázquez et al. 2017). A detailed introduction to these datasets can be found in `Appendix`.

Evaluation Metrics: Following previous works, we use AUROC for image-level anomaly classification and AUPR for pixel-level anomaly segmentation in our main paper. Here, we emphasize that AUPR is better for anomaly segmentation, where the imbalance issue is very extreme between normal and anomaly pixels (Zou et al. 2022). In `Appendix`, we also provide detailed comparisons using all metrics, including AUROC, AUPR, and $F1_{max}$.

Training and Testing Protocol: Following AnomalyCLIP (Zhou et al. 2024), we train AdaptCLIP using MVTec test data and evaluate zero-/few-shot performance on other datasets. For evaluation on MVTec, we train AdaptCLIP using VisA test data. For fair comparison, all models are trained and evaluated using the same protocol.

Competing Methods: We compare our AdaptCLIP with diverse state-of-the-art zero-/few-shot AD methods including zero-shot WinCLIP (Jeong et al. 2023), AnomalyCLIP (Zhou et al. 2024), AdaCLIP (Cao et al. 2024), and few-shot WinCLIP+ (Jeong et al. 2023), InCtrl (Zhu and Pang 2024) and AnomalyCLIP+. Here, AnomalyCLIP+ is a strong baseline we build on AnomalyCLIP (Zhou et al. 2024) by adding patch-level feature associations like WinCLIP+. More implementation details about AdaptCLIP and competing methods can be found in `Appendix`.

Comparisons with Zero-/Few-Shot Methods

Tabs. 1 and 2 present comparisons of AdaptCLIP to competing zero-/few-shot methods in image-level anomaly classification and pixel-level anomaly segmentation, respectively, on 8 real-world industrial and 4 medical AD datasets. Note that we only use image-level metrics to evaluate Br35H and Covid due to the lack of pixel-level annotations, and only report the results for Kvasir and Endo using pixel-level metrics since normal images are not included in these two datasets. Below, we analyze these results in detail.

Generalization on Industrial Domain: Generally, AdaptCLIP significantly outperforms all competing models on almost all industrial datasets across three few-shot settings, 1-shot, 2-shot and 4-shot. The performance of all methods generally gets better with more image prompts. Specifically, InCtrl (Zhu and Pang 2024) surpasses WinCLIP (Jeong et al. 2023) due to additional fine-tuning on a base training dataset. AnomalyCLIP (Zhou et al. 2024) further achieves better generalization, which verifies the importance of learning object-agnostic prompts. AdaptCLIP exhibits superior

Shots	Methods	Industrial									Medical		
		MVTec	VisA	BTAD	MVTec3D	DTD	KSDD	MPDD	Real-IAD	AVG	Kvasir	Endo	AVG
0	WinCLIP (Jeong et al. 2023)	18.2	5.4	12.9	5.3	9.8	7.1	14.1	3.3	9.5	27.8	23.8	25.8
	AdaCLIP† (Cao et al. 2024)	39.1	31.0	42.9	37.5	75.2	48.2	25.9	30.5	41.3	36.6	43.7	40.1
	AnomalyCLIP (Zhou et al. 2024)	34.5	21.3	45.5	30.5	62.6	51.9	28.9	26.7	37.7	39.6	46.6	43.1
	AdaptCLIP-Zero	38.3	26.1	41.8	31.4	68.7	58.3	25.3	28.2	39.7	45.3	52.0	48.7
1	WinCLIP+ (Jeong et al. 2023)	38.3±0.8	15.8±0.2	41.3±2.6	18.4±1.1	47.8±0.9	19.2±0.3	29.8±2.0	13.9±0.2	28.1	27.6±2.9	23.6±0.1	25.6
	InCtrl (Zhu and Pang 2024)	47.8±1.1	17.7±0.6	44.1±1.4	18.7±0.5	64.3±0.5	26.7±0.7	27.9±2.2	19.1±0.0	33.3	22.1±1.7	20.3±3.7	21.2
	AnomalyCLIP+ (Zhou et al. 2024)	40.8±0.1	24.8±0.9	41.3±1.1	30.6±1.1	67.4±0.4	47.5±0.5	34.2±0.8	27.9±0.0	39.3	46.9±3.9	47.8±4.9	47.4
	AdaptCLIP	53.7±0.9	38.9±0.3	60.6±1.0	40.7±0.6	76.9±0.1	57.8±1.2	33.5±2.5	36.6±0.1	49.8	49.2±4.7	52.4±4.7	50.8
2	WinCLIP+ (Jeong et al. 2023)	39.5±0.6	17.2±0.8	42.8±1.3	19.1±0.8	48.2±0.9	19.0±0.5	30.7±1.1	14.8±0.1	28.9	29.1±0.2	27.6±2.3	28.4
	InCtrl (Zhu and Pang 2024)	49.2±0.7	18.5±0.2	44.2±0.8	20.3±0.6	64.4±0.4	26.4±2.5	29.2±1.3	20.1±0.0	34.0	24.9±1.9	24.5±7.5	24.7
	AnomalyCLIP+ (Zhou et al. 2024)	41.5±0.1	26.2±0.7	41.9±0.6	32.4±1.5	68.1±0.2	47.6±0.4	35.3±1.1	28.1±0.0	40.1	47.3±2.9	49.6±4.8	48.5
	AdaptCLIP	55.1±0.5	40.7±0.6	61.0±0.6	42.3±1.1	77.4±0.2	57.5±1.1	35.0±0.7	37.8±0.1	50.9	49.0±4.1	53.1±4.2	51.1
4	WinCLIP+ (Jeong et al. 2023)	41.2±0.9	18.1±1.3	44.0±0.4	19.9±0.6	49.3±0.1	19.1±0.7	32.0±0.2	15.4±0.2	29.9	29.6±0.8	27.7±0.5	28.7
	InCtrl (Zhu and Pang 2024)	50.9±0.3	19.2±0.6	44.0±0.2	22.2±1.2	64.9±0.3	26.0±1.4	31.4±0.8	21.0±0.0	35.0	24.7±1.6	22.3±1.0	23.5
	AnomalyCLIP+ (Zhou et al. 2024)	42.4±0.0	27.5±1.1	45.8±3.0	33.4±1.3	68.5±0.2	46.4±0.7	36.8±1.0	28.2±0.0	41.1	45.9±1.5	49.2±3.4	47.6
	AdaptCLIP	57.2±0.8	41.8±0.6	62.3±0.3	44.5±0.3	78.2±0.2	56.4±1.4	37.4±1.1	39.1±0.3	52.1	47.5±2.7	52.2±3.1	49.9

Table 2: Pixel-level anomaly segmentation comparisons with AUPR metric on industrial and medical domains.

performance, outperforming AnomalyCLIP (Zhou et al. 2024) by a large margin (about 10%+ in pixel AUPR and 2%+ in image AUROC), particularly on challenging and large-scale datasets like VisA and Real-IAD. This reveals the power of comparative learning based on the joint contextual and aligned residual features for universal anomaly detection. Under zero-shot setting, AdaptCLIP-Zero significantly outperforms SOTA AdaCLIP (Cao et al. 2024) on anomaly classification, although it shows a slight weakness in industrial anomaly segmentation. However, AdaptCLIP is simpler, requires fewer learnable parameters (0.6M vs. 10.7M in Tab. 3), and generalizes better from the industrial to the medical domain. In addition, our one-shot AdaptCLIP easily outperforms zero-shot AdaCLIP (Cao et al. 2024) if only one normal image prompt is available.

Generalization on Medical Domain: Our AdaptCLIP performs strongly on medical AD regardless of zero-shot or few-shot settings when applying the same model trained on an industrial dataset (i.e., MVTec). Surprisingly, it significantly outperforms SOTA AdaCLIP on image anomaly classification (i.e., 6.3% in AUROC) and pixel anomaly segmentation (i.e., 8.6% in AUPR). Notably, our approach still works even when replacing normal image prompts with anomaly images. This is meaningful for some special datasets that don't contain any normal images, such as Kvasir and Endo. Here, this success is mainly due to the proposed spatial alignment mechanism, as well as a strong prior assumption that anomaly pixels are mostly sparse.

Efficiency Comparison: We measure complexity and efficiency by the number of parameters and the forward inference time, as shown in Tab. 3. The evaluation is performed on one V100 GPU with batch size 32. The number of parameters of AdaCLIP and AnomalyCLIP is 17 times and 9 times that of our AdaptCLIP, respectively. Compared to SOTA, AdaptCLIP achieves competitive inference time yet better AD performance. When extending from zero-shot to one-shot, AnomalyCLIP+ and our AdaptCLIP require almost no additional inference time, unlike earlier WinCLIP.

Shots	Methods	CLIP Models	Input Size	# Params (M)	Inf.Time (ms)
0	WinCLIP (Jeong et al. 2023)	ViT-B-16+240	240×240	208.4 + 0.0	201.3
		ViT-B-16+240	512×512	208.4 + 0.0	3912.6
	AdaCLIP (Cao et al. 2024)	ViT-L/14@336px	518×518	428.8 + 10.7	212.0
	AnomalyCLIP (Zhou et al. 2024)	ViT-L/14@336px	518×518	427.9 + 5.6	154.9
	AdaptCLIP-Zero	ViT-B-16+240	512×512	208.4 + 0.4	49.9
		ViT-L/14@336px	518×518	427.9 + 0.6	162.2
1	WinCLIP+ (Jeong et al. 2023)	ViT-B-16+240	240×240	208.4 + 0.0	339.5
		ViT-B-16+240	512×512	208.4 + 0.0	7434.9
	InCtrl (Zhu and Pang 2024)	ViT-B-16+240	240×240	208.4 + 0.3	337.0
	AnomalyCLIP+ (Zhou et al. 2024)	ViT-L/14@336px	518×518	427.9 + 5.6	158.6
	AdaptCLIP	ViT-B-16+240	512×512	208.4 + 1.4	54.0
		ViT-L/14@336px	518×518	427.9 + 1.8	168.2

Table 3: Complexity and efficiency comparisons.

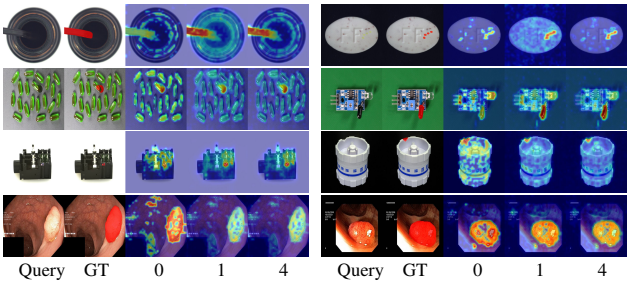

Figure 3: Qualitative comparisons of our AdaptCLIP with different prompt numbers on MVTec, VisA, Real-IAD, Kvasir and Endo. More qualitative results of AdaptCLIP can be found in `Appendix`. Best viewed in color and zoom in.

Qualitative Results: Fig. 3 shows some selected visualizations from industrial and medical testing images using AdaptCLIP. Generally, few-shot normal image prompts help AdaptCLIP segment anomalies more accurately and produce fewer false positives than in a zero-shot manner.

Comparisons with Few-/Full-Shot Methods

In Tab. 4, we compare few-shot AdaptCLIP with few-shot and full-shot unified AD models. It can be seen that AdaptCLIP is better than the early RegAD (Huang et al. 2022), and

Methods	Shots	MVTec	VisA	BTAD
AdaptCLIP	1	94.5 / 53.7	90.5 / 38.9	93.4 / 60.6
	4	96.6 / 57.2	93.1 / 41.8	93.3 / 62.3
MetaUAS (Gao 2024b)	1	90.7 / 59.3	81.2 / 42.7	-
PromptAD (Li et al. 2024b)	4	96.6 / 52.9	89.1 / 31.5	-
APRIL-GAN (Chen et al. 2023)	4	92.8 / 54.5	92.6 / 32.2	-
RegAD (Huang et al. 2022)	8	91.2 / 51.1	79.7 / 28.6	90.7 / 40.5
OneNIP (Gao 2024a)	full	97.9 / 63.7	92.5 / 43.3	92.6 / 56.8
SimpleNet (Liu et al. 2023b)	full	78.2 / 24.8	89.2 / 33.1	90.3 / 36.2
UniAD (You et al. 2022)	full	96.5 / 44.7	90.8 / 33.6	92.2 / 50.9

Table 4: Comparisons of anomaly classification and segmentation (AUROC/AUPR) with few- and full-shot methods.

No.	Methods	Shots	TA	VA	PQA	MVTec	VisA
0		0	✗	✗	✗	91.1 / 33.0	82.1 / 18.0
1	baselines	0	✓	✗	✗	92.2 / 31.4	82.9 / 19.7
2		0	✗	✓	✗	90.5 / 39.4	81.0 / 22.1
3	joint	0	✓	✓	✗	89.3 / 36.2	81.6 / 21.5
4	alternating	0	✓	✓	✗	93.5 / 38.3	84.8 / 26.1
5	w/o context	1	✗	✗	✓	62.6 / 7.0	85.3 / 28.7
6	w context	1	✗	✗	✓	88.1 / 50.2	88.9 / 38.1
7	AdaptCLIP	1	✓	✓	✓	94.2 / 52.5	92.0 / 38.8

Table 5: Ablation studies on different components.

comparable to MetaUAS (Gao 2024b) and PromptAD (Li et al. 2024b). It is worth noting that PromptAD (Li et al. 2024b) requires re-training with few-shot normal images while our method remains training-free on target domains. Furthermore, our method outperforms full-shot methods, such as SimpleNet (Liu et al. 2023b) and UniAD (You et al. 2022), and is also competitive with the latest OneNIP (Gao 2024a). In short, our method has shown excellent performance, especially in the open-world scenario for universal anomaly detection, although there is still some small gaps compared to SOTA full-shot methods.

Ablation Studies

To demonstrate the effectiveness of the proposed two key learning strategies, alternating learning, and comparative learning based on the joint contextual and aligned residual feature, and three adapters in AdaptCLIP, TA: Textual Adapter, VA: Visual Adapter, and PQA: Prompt-Query Adapter, we conduct experiments on MVTec and VisA, and report results in Tab. 5.

Simple but effective baselines. The first baseline is the naive CLIP (Line 0), and it is simple and effective for zero-shot anomaly detection only using two-class textual prompts. However, it is still weak in pixel-level anomaly segmentation. The individual textual adapter and visual adapter are two addtional baselines. Specifically, the textual adapter can be seen as an extreme simplification of AnomalyCLIP (Zhou et al. 2024), removing the textual prompt template and textual prompt tuning. The simple textual adapter performs better than the original AnomalyCLIP and naive CLIP in anomaly classification, although it is slightly inferior in anomaly segmentation (Lines 0 vs. 1). The visual adapter learns adaptive local patch tokens and global image tokens to align textual representations from CLIP in both patch and image levels. This significantly improves pixel-level anomaly segmentation (Lines 0 vs. 2).

Alternating learning is better than joint learning. We explore the impact of alternating learning and joint learning strategies on AdaptCLIP's performance. Alternating learning adapts visual or textual representations independently, while joint learning optimizes both representations simultaneously. As shown (Lines 3 vs. 4) in Tab. 5, the alternating learning strategy significantly enhances the performance of AdaptCLIP compared to joint learning. Alternating learning not only fully leverages the strong prior guidance of CLIP's visual and textual representations but also mitigates the risk of over-fitting due to fine-tuning on a small training dataset. Additionally, we observe that the visual adapter alone excels in anomaly segmentation (Line 2), whereas the textual adapter alone performs better in anomaly classification (Line 1). By integrating the alternating learning into visual and textual adapters, AdaptCLIP generally achieves superior anomaly detection performance (Line 4).

The joint of contextual information and aligned residual features performs better than residual features alone. The aligned residual feature captures the distinctions between anomalous features and their corresponding normal counterparts. It effectively eliminates features related to individual objects and may improve generalization. However, we realize that isolated residual features may lose contextual information about visual objects, resulting in degraded model performance or even training failure (Line 5). Therefore, we propose a joint feature learning based on both contextual and aligned residual features, which further significantly boosts the model's performance (Lines 6 vs. 5). This means contextual information is equally important for anomaly identification. Notably, the optimal performance for AdaptCLIP is achieved when all proposed components are integrated (Line 7).

Conclusion

In this paper, we introduce a universal anomaly detection task, which focuses on generalizing anomaly detection models across domains, such as industrial and medical, and in open scenarios, such as zero- or few-shot settings. Once the universal anomaly detection model is trained, it does not need any fine-tuning on the target dataset. Compared with single zero-shot or few-shot AD models, the universal anomaly detection model is more flexible, supporting zero-/few-shot inference via fixed or learnable textual prompts and a few normal image prompts, while providing both image-level and pixel-level anomaly predictions. We propose a universal anomaly detection framework, AdaptCLIP, which alternately learns adaptive visual representations and text prompt embeddings, as well as jointly learns comparisons based on the contextual information of query image and the aligned residual features between the query and the prompt. Extensive experiments on 8 standard industrial and 4 medical datasets show that AdaptCLIP significantly outperforms current competitive models in multiple settings.

References

Aota, T.; Tong, L. T. T.; and Okatani, T. 2023. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *WACV*.

Bengio, Y.; Yao, L.; Alain, G.; and Vincent, P. 2013. Generalized denoising auto-encoders as generative models. In *NeurIPS*.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec-AD: A comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2020. Uninformed Students: Student-teacher anomaly detection with discriminative latent embeddings. In *CVPR*.

Bergmann., P.; Jin., X.; Sattlegger., D.; and Steger., C. 2022. The MVTec 3D-AD Dataset for Unsupervised 3D Anomaly Detection and Localization. In *VISAPP*.

Cao, Y.; Zhang, J.; Frittoli, L.; Cheng, Y.; Shen, W.; and Boracchi, G. 2024. AdaCLIP: Adapting CLIP with Hybrid Learnable Prompts for Zero-Shot Anomaly Detection. In *ECCV*.

Chen, X.; Han, Y.; and Zhang, J. 2023. A zero-/few-shot anomaly classification and segmentation method for CVPR 2023 (VAND) workshop challenge tracks 1 &2. In *CVPRW*.

Defard, T.; Setkov, A.; Loesch, A.; and Audigier, R. 2021. PaDiM: a patch distribution modeling framework for anomaly detection and localization. In *ICPR*.

Deng, H.; and Li, X. 2022. Anomaly detection via reverse distillation from one-class embedding. In *CVPR*.

Ding, C.; Pang, G.; and Shen, C. 2022. Catching both gray and black swans: Open-set supervised anomaly detection. In *CVPR*.

Fang, Z.; Wang, X.; Li, H.; Liu, J.; Hu, Q.; and Xiao, J. 2023. FastRecon: Few-shot Industrial Anomaly Detection via Fast Feature Reconstruction. In *ICCV*.

Gao, B.-B. 2024a. Learning to Detect Multi-class Anomalies with Just One Normal Image Prompt. In *ECCV*.

Gao, B.-B. 2024b. MetaUAS: Universal Anomaly Segmentation with One-Prompt Meta-Learning. In *NeurIPS*.

Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M. R.; Venkatesh, S.; and Hengel, A. v. d. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *ICCV*.

Gui, G.; Gao, B.-B.; Liu, J.; Wang, C.; and Wu, Y. 2024. Few-shot anomaly-driven generation for anomaly classification and segmentation. In *ECCV*.

Hamada, A. 2020. Br35h: Brain tumor detection 2020.

Hou, J.; Zhang, Y.; Zhong, Q.; Xie, D.; Pu, S.; and Zhou, H. 2021. Divide-and-Assemble: Learning block-wise memory for unsupervised anomaly detection. In *ICCV*.

Huang, C.; Guan, H.; Jiang, A.; Zhang, Y.; Spratling, M.; and Wang, Y.-F. 2022. Registration based few-shot anomaly detection. In *ECCV*.

Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; and Dabeer, O. 2023. WinCLIP: Zero-/few-shot anomaly classification and segmentation. In *CVPR*.

Jezek, S.; Jonak, M.; Burget, R.; Dvorak, P.; and Skotak, M. 2021. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *ICUMT*.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; de Lange, T.; Johansen, D.; and Johansen, H. D. 2020. Kvasir-seg: A segmented polyp dataset. In *MMM*.

Lei, J.; Hu, X.; Wang, Y.; and Liu, D. 2023. PyramidFlow: High-Resolution Defect Contrastive Localization using Pyramid Normalizing Flow. In *CVPR*.

Li, A.; Qiu, C.; Kloft, M.; Smyth, P.; Rudolph, M.; and Mandt, S. 2023. Zero-Shot Anomaly Detection via Batch Normalization. In *NeurIPS*.

Li, C.-L.; Sohn, K.; Yoon, J.; and Pfister, T. 2021. CutPaste: Self-supervised learning for anomaly detection and localization. In *CVPR*.

Li, X.; Huang, Z.; Xue, F.; and Zhou, Y. 2024a. MuSc: Zero-Shot Industrial Anomaly Classification and Segmentation with Mutual Scoring of the Unlabeled Images. In *ICLR*.

Li, X.; Zhang, Z.; Tan, X.; Chen, C.; Qu, Y.; Xie, Y.; and Ma, L. 2024b. PromptAD: Learning Prompts with only Normal Samples for Few-Shot Anomaly Detection. In *CVPR*.

Liu, W.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2023a. Diversity-measurable anomaly detection. In *CVPR*.

Liu, Z.; Zhou, Y.; Xu, Y.; and Wang, Z. 2023b. SimpleNet: A simple network for image anomaly detection and localization. In *CVPR*.

Mishra, P.; Verk, R.; Fornasier, D.; Piciarelli, C.; and Foresti, G. L. 2021. VT-ADL: A vision transformer network for image anomaly detection and localization. In *ISIE*.

Perera, P.; Nallapati, R.; and Xiang, B. 2019. OCGAN: One-class novelty detection using GANs with constrained latent representations. In *CVPR*.

Qu, Z.; Tao, X.; Prasad, M.; Shen, F.; Zhang, Z.; Gong, X.; and Ding, G. 2024. VCP-CLIP: A visual context prompting model for zero-shot anomaly segmentation. In *ECCV*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Rahman, T.; Khandakar, A.; Qiblawey, Y.; Tahir, A.; Kiranyaz, S.; Kashem, S. B. A.; Islam, M. T.; Al Maadeed, S.; Zughaier, S. M.; Khan, M. S.; et al. 2021. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Computers in biology and medicine*.

Rao, R. P.; and Ballard, D. H. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1).

Rippel, O.; Mertens, P.; and Merhof, D. 2021. Modeling the distribution of normal data in pretrained deep features for anomaly detection. In *ICPR*.

Ristea, N.-C.; Madan, N.; Ionescu, R. T.; Nasrollahi, K.; Khan, F. S.; Moeslund, T. B.; and Shah, M. 2022. Self-supervised predictive convolutional attentive block for anomaly detection. In *CVPR*.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2022. Towards total recall in industrial anomaly detection. In *CVPR*.

Rudolph, M.; Wehrbein, T.; Rosenhahn, B.; and Wandt, B. 2022. Fully convolutional cross-scale-flows for image-based defect detection. In *WACV*.

Salehi, M.; Sadjadi, N.; Baselizadeh, S.; Rohban, M. H.; and Rabiee, H. R. 2021. Multiresolution knowledge distillation for anomaly detection. In *CVPR*.

Sheynin, S.; Benaim, S.; and Wolf, L. 2021. A hierarchical transformation-discriminating generative model for few shot anomaly detection. In *ICCV*.

Tabernik, D.; Šela, S.; Skvarč, J.; and Skočaj, D. 2020. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3).

Vázquez, D.; Bernal, J.; Sánchez, F. J.; Fernández-Esparrach, G.; López, A. M.; Romero, A.; Drozdzal, M.; Courville, A.; et al. 2017. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and composing robust features with denoising autoencoders. In *ICML*.

Wang, C.; Zhu, W.; Gao, B.-B.; Gan, Z.; Zhang, J.; Gu, Z.; Qian, S.; Chen, M.; and Ma, L. 2024. Real-IAD: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *CVPR*.

Wang, G.; Han, S.; Ding, E.; and Huang, D. 2021a. Student-Teacher Feature Pyramid Matching for Anomaly Detection. *BMVC*.

Wang, S.; Wu, L.; Cui, L.; and Shen, Y. 2021b. Glancing at the patch: Anomaly localization with global and local feature comparison. In *CVPR*.

Xie, G.; Wang, J.; Liu, J.; Zheng, F.; and Jin, Y. 2023. Pushing the limits of fewshot anomaly detection in industry vision: Graphcore. In *ICLR*.

Yan, X.; Zhang, H.; Xu, X.; Hu, X.; and Heng, P.-A. 2021. Learning semantic context from normal samples for unsupervised anomaly detection. In *AAAI*.

Yao, X.; Li, R.; Qian, Z.; Luo, Y.; and Zhang, C. 2023a. Focus the Discrepancy: Intra-and inter-correlation learning for image anomaly detection. In *ICCV*.

Yao, X.; Li, R.; Zhang, J.; Sun, J.; and Zhang, C. 2023b. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *CVPR*.

You, Z.; Cui, L.; Shen, Y.; Yang, K.; Lu, X.; Zheng, Y.; and Le, X. 2022. A unified model for multi-class anomaly detection. In *NeurIPS*.

Zaheer, M. Z.; Lee, J.-h.; Astrid, M.; and Lee, S.-I. 2020. Old is Gold: Redefining the adversarially learned one-class classifier training paradigm. In *CVPR*.

Zavrtanik, V.; Kristan, M.; and Skočaj, D. 2021a. DRAEM: A discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*.

Zavrtanik, V.; Kristan, M.; and Skočaj, D. 2021b. Reconstruction by inpainting for visual anomaly detection. *PR*, 112.

Zhao, Y. 2023. OmniAL: A unified CNN framework for unsupervised anomaly localization. In *CVPR*.

Zhou, Q.; Pang, G.; Tian, Y.; He, S.; and Chen, J. 2024. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *ICLR*.

Zhu, J.; and Pang, G. 2024. Toward Generalist Anomaly Detection via In-context Residual Learning with Few-shot Sample Prompts. In *CVPR*.

Zou, Y.; Jeong, J.; Pemula, L.; Zhang, D.; and Dabeer, O. 2022. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *ECCV*.