

AA-CLIP: Enhancing Zero-Shot Anomaly Detection via Anomaly-Aware CLIP

Wenxin Ma^{1,2} Xu Zhang^{1,2} Qingsong Yao⁵ Fenghe Tang^{1,2} Chenxu Wu^{1,2}
 Yingtai Li^{1,2} Rui Yan^{1,2} Zihang Jiang^{1,2,*} S.Kevin Zhou^{1,2,3,4*}

¹ School of Biomedical Engineering, Division of Life Sciences and Medicine, USTC

² MIRACLE Center, Suzhou Institute for Advance Research, USTC

³ Key Laboratory of Intelligent Information Processing of CAS, ICT, CAS

⁴ State Key Laboratory of Precision and Intelligent Chemistry, USTC

⁵ Stanford University

wxma@mail.ustc.edu.cn jzh0103@ustc.edu.cn s.kevin.zhou@gmail.com

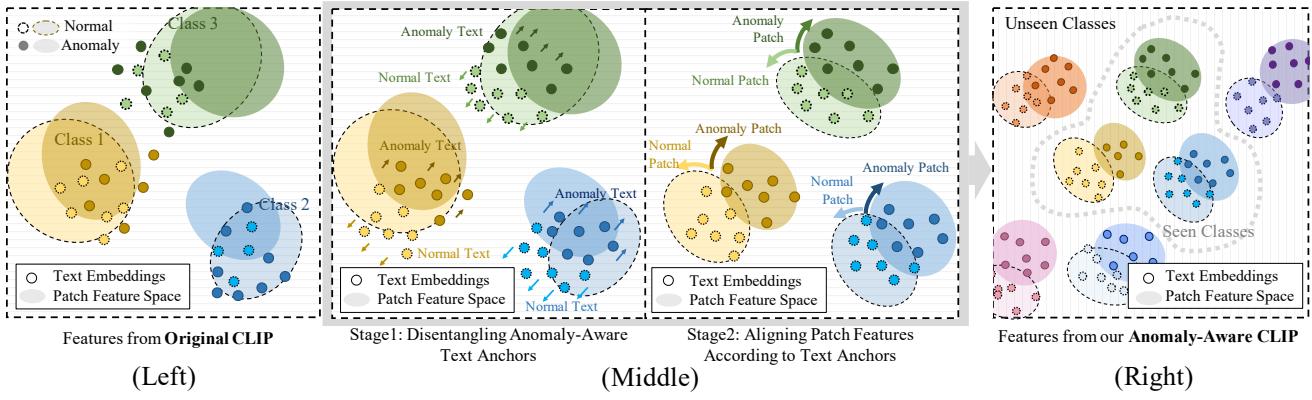


Figure 1. **(Left) CLIP’s anomaly unawareness:** Category-level image-text alignment in pre-training leads to CLIP’s vague distinctions in anomaly/normal semantics and inaccurate patch-text alignment. **(Middle) Our two-stage adaptation strategy:** In Stage1, anomaly and normal text features are disentangled as anchors in text space; in Stage2, patch-level visual features are trained to align to these anchors, forming Anomaly-Aware CLIP. **(Right) Generalizable anomaly awareness:** Our method enables CLIP with generalizable anomaly awareness for both known and unseen classes.

Abstract

Anomaly detection (AD) identifies outliers for applications like defect and lesion detection. While CLIP shows promise for zero-shot AD tasks due to its strong generalization capabilities, its inherent **Anomaly-Unawareness** leads to limited discrimination between normal and abnormal features. To address this problem, we propose **Anomaly-Aware CLIP** (AA-CLIP), which enhances CLIP’s anomaly discrimination ability in both text and visual spaces while preserving its generalization capability. AA-CLIP is achieved through a straightforward yet effective two-stage approach: it first creates anomaly-aware text anchors to differentiate normal and abnormal semantics clearly, then aligns patch-level visual features with these anchors for precise anomaly localization. This two-stage strategy, with the help of residual adapters, gradually adapts CLIP in a controlled man-

ner, achieving effective AD while maintaining CLIP’s class knowledge. Extensive experiments validate AA-CLIP as a resource-efficient solution for zero-shot AD tasks, achieving state-of-the-art results in industrial and medical applications. The code is available at <https://github.com/Mwxinnn/AA-CLIP>.

1. Introduction

Anomaly detection (AD) involves modeling the distribution of a dataset to identify outliers, such as defects in industrial products [2] or lesions in medical images [12]. Despite that previous AD frameworks [9, 10, 14, 22, 30, 56] effectively detect anomalies when sufficient labeled data is available for specific classes, their high resource demands often limit their generalization ability to novel and rare classes. This limitation is particularly challenging in real-world scenarios where collecting comprehensive labeled datasets for AD is often infeasible, necessitating the exploration of low-shot

*Corresponding author.

learning and transfer learning approaches.

Contrastive Language-Image Pretraining (CLIP) model has emerged as a promising solution, demonstrating remarkable generalization capabilities across various zero-shot tasks [23–25, 41]. Building upon CLIP’s success, several recent studies have adapted CLIP for few/zero-shot AD tasks by utilizing anomaly-related descriptions to guide the detection of anomalous regions. Specifically, the vision encoder is trained to map anomaly images to visual features that align more closely with text features of abnormal descriptions than with those of normal descriptions [28, 29, 48, 58]. Further works [5, 6, 16, 40] have focused on enhancing CLIP’s patch-level feature representations to achieve better alignment with text features, resulting in improved anomaly localization performance.

These methods depend on text features that need to be anomaly-aware to effectively differentiate abnormalities. However, recent studies highlight CLIP’s limitations in fine-grained semantic perception and reasoning [20, 21, 35, 37, 44, 45]. Upon exploring CLIP’s texture features for AD, we observe that while CLIP’s text encoder effectively captures object-level information, it struggles to reliably distinguish between normal and abnormal semantics. As shown in conceptual visualization Fig. 1(left) and sampled examples in Fig. 2, CLIP has the intrinsic **Anomaly-Unawareness** problem: the overlap of normal and abnormal texture features hampers the precision of text-guided anomaly detection. We argue that making CLIP anomaly-aware — by establishing clearer distinctions between normal and abnormal semantics in the text space — is essential for guiding the vision encoder to precisely detect and localize anomalies.

This observation drives us to improve CLIP-based zero-shot AD through enhancing anomaly discrimination in text space, achieved with our method **Anomaly-Aware CLIP (AA-CLIP)** — a CLIP model with anomaly-aware information encoded. AA-CLIP is implemented through a novel two-stage adaptation approach. In the first stage, AA-CLIP adapts the text encoder with frozen visual encoder, creating “anchors” for anomaly-aware semantics within the text space for each trained class. As illustrated in Fig. 1(middle), each class’s text features are disentangled to distinct anchors, with clear abnormality discrimination. Notably, this disentanglement also applies to novel, unseen classes, supporting effective zero-shot inference in AD tasks (refer to Fig. 1(right)). In the second stage, AA-CLIP aligns patch-level visual features with these specially adapted texture anchors, guiding CLIP’s visual encoder to concentrate on anomaly-relevant regions. This two-stage approach ensures a focused and precise anomaly detection framework.

Importantly, as CLIP is extensively trained on massive data, to preserve its pre-trained knowledge, we utilize simple-structured Residual Adapters in both stages. This design enables a controlled adaptation of CLIP while en-

hancing its capability to handle fine-grained AD tasks without sacrificing its generalization ability.

Our extensive experiments in both industrial and medical domains demonstrate that our straightforward approach equips CLIP with improved zero-shot AD ability, even in data-limited scenarios. By training with a minimal sample — such as one normal sample and one anomaly sample (2-shot) per class — and testing across unseen datasets, our method achieves zero-shot performance comparable to other CLIP-based AD techniques. With only 64-shot of each class seen in the training set, our method reaches state-of-the-art (SOTA) results in cross-dataset zero-shot testing, validating our method’s ability to maximize the CLIP’s potential for AD with a minimal data requirement.

Our contributions are summarized as follows:

1. *Anomaly-Aware CLIP with enhanced and generalizable anomaly-discriminative ability.* We introduce AA-CLIP which is more sensitive to anomalies sequentially in text and visual spaces, encoding anomaly-aware information into the original CLIP.
2. *Efficient adaptation using residual adapters.* We implement simple residual adapters to boost zero-shot anomaly detection performance without compromising the model’s generalization ability.
3. *SOTA performance with high training efficiency.* Our method achieves SOTA results across diverse datasets, showing robust anomaly detection capabilities even with limited training samples.

2. Related Work

Traditional Anomaly Detection in images involves modeling the normal data distribution to detect rare and diverse unexpected signals within visual data [43, 51, 59]. Reconstruction-based [10, 15, 31, 32, 52, 54], augmentation-based [30, 43, 47, 53, 56] and discriminative [9, 14, 22, 30, 42, 59] methods are typically used to facilitate better modeling. Despite the huge progress of traditional anomaly detection methods, their effectiveness relies heavily on a well-modeled normal data distribution. Without sufficient normal data, their ability to accurately detect anomalies is significantly reduced.

CLIP, trained on a vast amount of image-text data, leverages contrastive learning alongside powerful language models and visual feature encoders to capture robust concepts. This combination enables CLIP to achieve impressive zero-shot performance on image classification, as it can generalize well to new categories without requiring task-specific training [23–26, 41, 50]. More recently, numerous studies [8, 11, 34, 38] have explored ways to transfer the knowledge embedded in CLIP models to a variety of downstream tasks, yielding promising results in fields like image captioning, image-text retrieval, and image generation. These efforts demonstrate CLIP’s versatility and potential to drive

advancements across diverse applications.

Despite the rapid advancements achieved by CLIP, numerous studies have highlighted persistent limitations in the features it extracts. While CLIP demonstrates strong generalization across various tasks, it often struggles to capture nuanced details and essential spatial relationships, which are crucial for tasks demanding precise boundary delineation and fine-grained feature extraction. This limitation results in suboptimal performance in downstream applications, especially those that require high levels of detail, such as object detection, scene segmentation, or tasks in medical imaging [13, 28, 29, 36, 48, 49, 55, 58]. As a result, leveraging CLIP for fine-granular tasks frequently necessitates task-specific adaptations to bridge the gap between its generalized feature extraction and the precision required for specialized applications.

CLIP-based Anomaly Detection There have been several efforts to leverage CLIP for AD tasks. One of the pioneering approaches, WinCLIP [18], proposes a method for extracting and aggregating visual features from multiple levels to align with text features, demonstrating the potential of CLIP in this context. Subsequent research investigates various adaptation methods to bridge the gap between natural domains and the AD domain, resulting in performance improvements. For instance, [6, 7, 17] focus on refining visual features by employing adapters to enhance patch-level visual representations. However, these approaches often rely on text embeddings from the original CLIP model as soft supervision and overlook a critical limitation of CLIP in AD: its uncleanness in distinguishing between anomalous and normal semantics, particularly within the text encoder, resulting in suboptimal performance. Other works have employed prompt-learning-based methods [5, 40, 57], introducing learnable embeddings into the text encoder to better represent abnormality. However, the class information in CLIP can be damaged, potentially degrading generalization, especially in data-limited and zero-shot settings.

Different from previous methods, we are the first to investigate CLIP’s inherent limitation in capturing anomaly-aware information, specifically in differentiating between normal and anomalous semantics in text prompts. Rather than relying solely on the original anomaly-unaware text embeddings or unaltered feature spaces, our method is able to refine the embeddings to actively incorporate anomaly-discriminative representations.

3. Method

3.1. Overview

3.1.1. Problem Formulation

Zero-shot AD models are trained to identify anomalous samples whose categories may be unseen in the training dataset. Specifically, the model is expected to learn

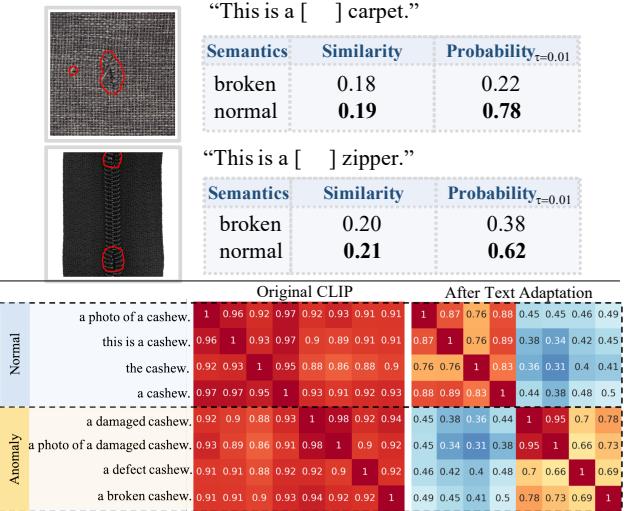


Figure 2. **(Top)** Examples illustrating CLIP’s Anomaly Unawareness. Despite the obvious anomalies present in the images, image features have higher similarities to normal descriptions, rather than anomaly descriptions, mistakenly. This problem is enlarged with a low temperature τ . **(Bottom)** Text Feature Similarity Heatmap among Normal and Anomaly Descriptions: Original CLIP vs. After Text Adaptation. Red indicates high similarity. In original CLIP, normal features exhibit strong similarity with anomaly features, whereas text adaptation successfully separates them, clarifying the semantic distinctions between normal and anomaly descriptions.

both normal and abnormal patterns that are shared across different classes given a training set \mathcal{D}_{train} with normal or anomalous samples, in order to be capable of performing AD tasks on a series of different test datasets $\{\mathcal{D}_{test}^1, \mathcal{D}_{test}^2, \dots, \mathcal{D}_{test}^n\}$, where each \mathcal{D}_{test}^i is distinct from \mathcal{D}_{train} . Image-level AD can be formally defined as a binary classification problem, where the model aims to classify samples $x \in \mathcal{D}$ as either normal ($y = 0$) or anomalous ($y = 1$). Anomaly segmentation extends this concept to pixel-level with mask S , aiming to identify anomalous regions by highlighting pixels associated with anomalies.

3.1.2. Current Challenges

Anomaly Unawareness in CLIP: The CLIP-based AD method classifies visual features as “anomalies” if they exhibit greater similarity to anomaly prompt embeddings than to normal prompt embeddings, thus requiring well-defined boundaries between these two kinds of prompts. However, in real applications, CLIP’s text embeddings often lack the clear separability needed to reliably distinguish between normal and anomalous classes.

We observe that, despite the visible defects in example images from the MVTec-AD [2], their features exhibit higher cosine similarity with “normal” prompts than with correct “anomaly” descriptions (see Fig. 2 (top)), indicating CLIP’s inaccurate semantic understanding. Without adap-

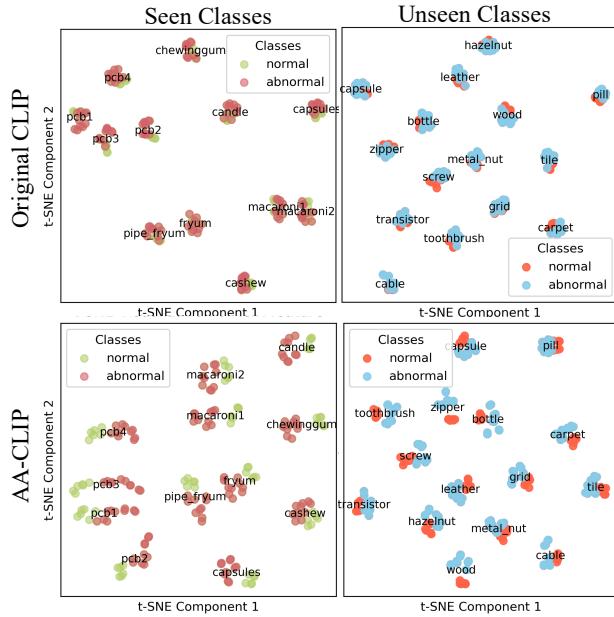


Figure 3. **t-SNE Visualization of Text Features from Original CLIP vs. AA-CLIP.** Each point represents a text feature encoded from a prompt. Original CLIP’s normal and anomaly text features are intertwined, while our method effectively disentangles them. This disentanglement is generalizable to novel classes, validating the anomaly-awareness of our model.

tation, there persists a high similarity between the normal and abnormal text embeddings of a single class, as shown in Fig. 2 (bottom), suggesting a potential entanglement of normal and anomaly semantics within text space. We term this limitation **Anomaly Unawareness** and attribute it to the training process of CLIP: it is primarily trained on general, non-anomalous datasets and lacks specific guidance on defect detection. Consequently, it is challenging to rely on original CLIP embeddings to detect subtle or context-specific anomalies.

This issue remains evident across different categories in our t-SNE analysis: as shown in Fig. 3 (top), only subtle separations are observed within an object cluster, where text embeddings for both normal and abnormal semantics are intermixed. This entangled pattern may potentially lead to anomaly-unaware text-image alignment, which reinforces the necessity to *adapt CLIP’s to enhance its ability of anomaly-awareness*.

Embedding Adaptation Dilemma: Discussion above renders the adaptation of CLIP essential for effective AD. However, since CLIP’s embeddings are already optimized through extensive pretraining, it could be susceptible to overfitting to new dataset during adaptation. Overfitting convergence leads to minimized intra-class distinctions in the training data, often at the expense of the feature separability for effective generalization to unseen data.

To address this, a *carefully controlled refinement* is crucial to preserve CLIP’s generalization capabilities while enhancing its sensitivity to anomalies.

3.1.3. Overview of Our Solution

Motivated by Sec. 3.1.2, we propose **Anomaly-Aware CLIP (AA-CLIP)** with improved anomaly awareness. As shown in Fig. 4, AA-CLIP is trained through a two-stage training strategy that sequentially adapts the semantic-rich text space and detail-focused visual space, with original CLIP parameters remaining frozen. In the first stage (see Fig. 4 (Top)), we incorporate Residual Adapters into the shallow layers of the text encoder, and the visual features from the fixed image encoder serve as a stable reference for optimization. A Disentangle Loss is purposed to enforce effective discrimination by ensuring independence between normal and anomaly embeddings. In the second stage, we integrate Residual Adapters into the shallow layers of the visual encoder to align patch-level features with the fixed, specially adapted texture features from the fixed text encoder (see in Fig. 4 (Bottom)). Ultimately, our AA-CLIP succeeds in equipping CLIP with anomaly awareness across seen and unseen classes, as shown in Fig. 3 (bottom).

3.2. AA-CLIP with Two-Stage Adaptation Strategy

3.2.1. Residual Adapter

To preserve CLIP’s pre-trained knowledge while enabling targeted adaptation, we introduce lightweight Residual Adapters in the shallow layers (up to layer K) of both text and vision encoders.

The output feature $x^i \in \mathbb{R}^{N \times d}$ of CLIP’s i -th ($i \leq K$) transformer layer is fed into the i -th adapter, outputting adapted feature x_{residual}^i , as shown in Eq. (1),

$$x_{\text{residual}}^i = \text{Norm}(\text{Act}(W^i x^i)), \quad (1)$$

where $W^i \in \mathbb{R}^{d \times d}$ is the trainable linear weight of i -th adapter, $\text{Act}(\cdot)$ is an activation function, and $\text{Norm}(\cdot)$ is a normalizing function. The original feature x^i and the enhanced feature x_{residual}^i are fused in a weighted manner, generating x_{enhanced}^i , the input to the next transformer layer, as shown in Eq. (2),

$$x_{\text{enhanced}}^i = \lambda x_{\text{residual}}^i + (1 - \lambda) x^i, \quad (2)$$

where λ is a hyper-parameter to control the residual ratio, adjusting the fusing degree of AD-specific knowledge for preserving the original CLIP’s generalization ability and improved performance.

3.2.2. Two-Stage Training Strategy

Disentangling Anomaly-Aware Text Anchors: In the first stage, our objective is to learn anomaly-discriminative text anchors by adapting the text encoder while keeping the

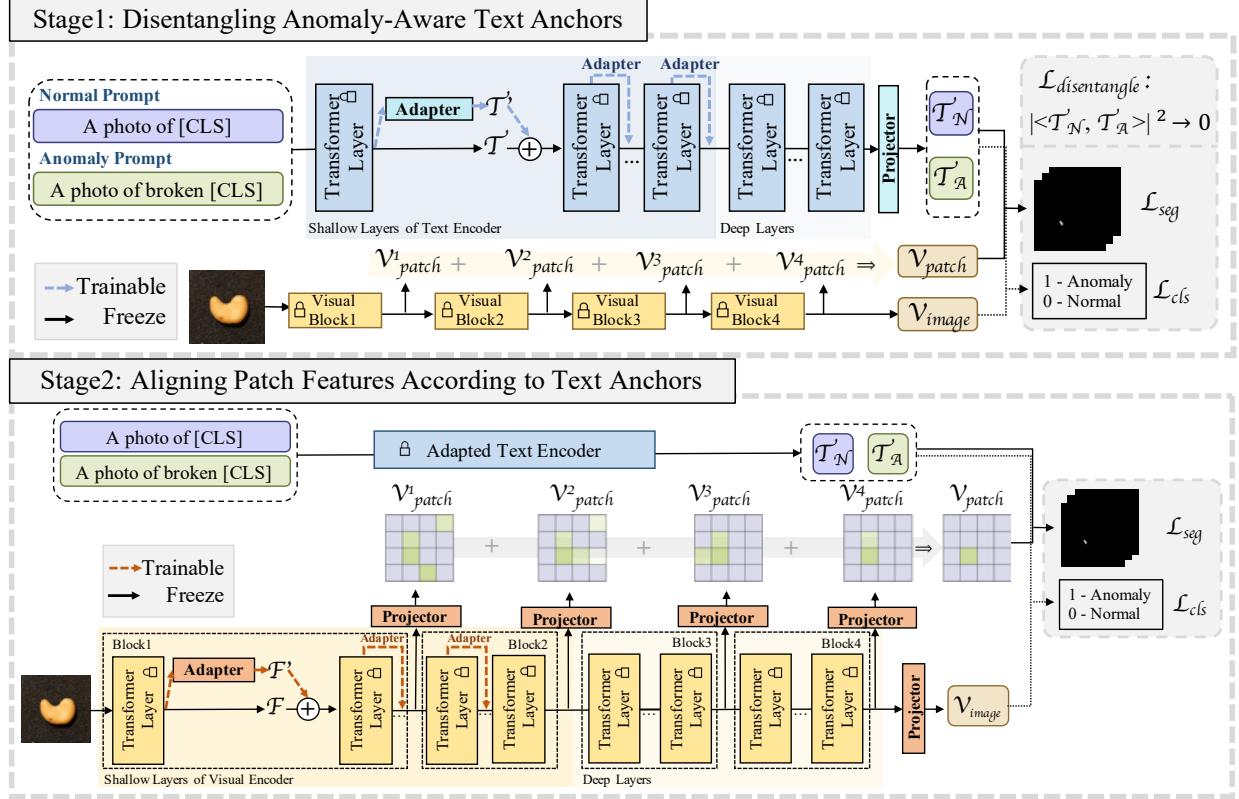


Figure 4. **The Two-Stage Training Pipeline of Anomaly-Aware CLIP.** In the first stage, the text encoder of AA-CLIP is trained to identify anomaly-related semantics, helped by a disentangle loss. In the second stage, patch features are aligned with these text anchors. Both stages are achieved by the integration of Residual Adapters into the shallow layers of CLIP’s backbone. This controlled adaptation enables CLIP to effectively distinguish anomalies, which forms our Anomaly-Aware CLIP.

image encoder fixed. We incorporate Residual Adapters into the first K_T layers of the CLIP text encoder, as illustrated in Fig. 4 (Top), and set the final projector in the text encoder to be learnable to facilitate improved alignment.

Using prompts designed to encapsulate both normal and anomalous semantics (as detailed in Appendix), text encoder generates corresponding high-level embeddings. The average embeddings of the normal and anomaly prompts serve as our initial text anchors, denoted as T_N and $T_A \in \mathbb{R}^d$, respectively. These anchors are refined by being aligned with visual features extracted from an enhanced CLIP visual encoder, as [27, 57]. Alignment is conducted at both image and patch levels to incorporate both global and local semantics. By calculating the cosine similarity between these anchors and the image features $V_{image} \in \mathbb{R}^d$ or patch features $V_{patch} \in \mathbb{R}^{N \times d}$, as shown in Eq. (3),

$$p_{cls} = \text{CosSim}(V_{image}, [T_N, T_A]), \quad (3)$$

$$p_{seg}^o = \text{CosSim}(V_{patch}, [T_N, T_A]),$$

where $[\cdot, \cdot]$ means concatenate operation, we obtain the classification prediction $p_{cls} \in \mathbb{R}^2$ and the segmentation prediction $p_{seg}^o \in \mathbb{R}^{N \times 2}$. The segmentation prediction p_{seg}^o is then reshaped and upsampled to $p_{seg} \in \mathbb{R}^{H \times W \times 2}$ to align

with the height H and width W of segmentation mask S . Following previous works [5, 6, 17, 57], we compute the classification loss \mathcal{L}_{cls} and segmentation loss \mathcal{L}_{seg} to optimize parameters, as specified in Eq. (4). Specifically, the classification loss is a binary cross-entropy that compares classification predictions with ground-truth labels y , and the segmentation loss is a combination of dice loss and focal loss applied to segmentation predictions and the anomaly segmentation mask S .

$$\begin{aligned} \mathcal{L}_{cls} &= \text{BCE}(p_{cls}, y), \\ \mathcal{L}_{seg} &= \text{Dice}(p_{seg}, S) + \text{Focal}(p_{seg}, S), \\ \mathcal{L}_{align} &= \mathcal{L}_{cls} + \mathcal{L}_{seg}. \end{aligned} \quad (4)$$

To enhance the separation between normal and anomaly text embeddings, we introduce a Disentangle Loss encouraging orthogonality between T_N and T_A to minimize correlation, as in Eq. (5):

$$\mathcal{L}_{dis} = | \langle T_N, T_A \rangle |^2. \quad (5)$$

The Disentangle Loss \mathcal{L}_{dis} is incorporated into the alignment loss \mathcal{L}_{align} as a regularization term, weighted by a factor γ , which forms the total loss, as in Eq. (6):

$$\mathcal{L}_{total} = \mathcal{L}_{align} + \gamma \mathcal{L}_{dis}. \quad (6)$$

In this stage, the distinction between normal and anomaly semantics is embedded into CLIP’s text encoder while its original object-recognition capability is preserved. Figure 3 indicates that this ability of anomaly-awareness is robust and generalizable to novel classes.

Aligning Patch Features According to Text Anchors: Anomaly-aware semantic anchors can facilitate the adaptation of patch features, thereby improving the effectiveness and generalizability of anomaly localization. To achieve alignment between patch features and anchors from the previous stage, we introduce trainable Residual Adapters within the initial K_I layers of the CLIP visual encoder.

Features with multi-granularities are utilized to enhance segmentation [6, 18, 57]. Specifically, as shown in Fig. 4 (bottom), the intermediate output feature F^i are extracted from four distinct granularities. These multi-granularity features are then projected to align with the channel of text anchors via a trainable projector $Proj_i(\cdot)$, yielding V_{patch}^i at four distinct levels of granularity. The aggregated output V_{patch} is computed by summing individual V_{patch}^i outputs, as in Eq. (7):

$$V_{patch}^i = Proj_i(F^i), i \in \{1, 2, 3, 4\}$$

$$V_{patch} = \sum_{i=1}^4 V_{patch}^i. \quad (7)$$

The cosine similarity scores between the aggregated V_{patch} and the text anchors are calculated to generate patch-level predictions as Eq. (3), resulting in the prediction maps.

During training, alignment is guided by the loss function defined in Eq. (4), facilitating both global and local alignment. During inference, anomaly prediction maps and corresponding anomaly scores are derived by comparing the similarity scores of visual features against normal and anomaly text embeddings.

4. Experiments

4.1. Experiment Setups

Datasets We evaluate our model on 11 widely used benchmarks, as previous AD works [5, 6, 17, 18, 57], with distinct foreground objects spanning a variety of modalities, including photography, endoscopy, CT, MRI, and OCT. For the industrial domain, we use MVTec AD [2], VisA [60], BTAD [33] and MPDD [19]. For medical domain, we use brain MRI, liver CT and retina OCT from BMAD [1], and four different colon polyp detection datasets with different views (CVC-ClinicDB [4], CVC-ColonDB [3], Kvasi-SEG [39] and CVC-300 [46]). Each dataset has both image-level labels and pixel-level masks for evaluation.

We train our model on a real-world industrial AD dataset - VisA [60] - in which objects are different from other datasets. Results of VisA are obtained using MVTec-AD as

the training dataset. To demonstrate adaptation efficiency, we conduct training under various data levels: 2-shot per class, 16-shot per class, 64-shot per class, and full-shot. The corresponding number of samples are randomly selected from each class, while maintaining a consistent 1:1 ratio between normal and anomaly samples.

Metrics Following [5, 6, 9, 10, 17, 18, 42, 53], we use the Area Under the Receiver Operating Characteristic Curve (AUROC) as the metric. We compute AUROC at both the image and pixel levels to comprehensively assess the model’s effectiveness in detecting and localizing anomalies.

Implementation Details Following [5, 6, 40, 57], we use OpenCLIP with the ViT-L/14 architecture as the backbone, and input images are resized to 518×518 . All parameters of CLIP remain frozen. We set λ to 0.1, K_T to 3, K_I to 6, and γ to 0.1. For multi-level feature extraction, we utilize outputs from the 6-th, 12-th, 18-th, and 24-th layers of the visual encoder to compose the overall output. For the first stage, we train the model for 5 epochs with a learning rate of 1×10^{-5} . For the second stage, we continue training for 20 epochs, adjusting the learning rate to 5×10^{-4} . Parameters are updated by Adam optimizers. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU. More details are available in Appendix.

4.2. Comparison with SOTA Methods

We compare our method against CLIP and several recent SOTA models. Among them, WinCLIP [18], VAND [6] and MVFA-AD [17] use original CLIP text encoder, and AnomalyCLIP [57] and AdaCLIP [5] incorporate learnable prompts. To ensure a fair comparison, we re-train models that are originally trained on different datasets to match the dataset settings of other approaches (detailed in Appendix).

Quantitative results are presented in Tab. 1 and Tab. 2. Although adapting only the patch feature with original text embeddings has made progress in AD, the superior performance of AA-CLIP highlights its effective disentanglement of anomaly-discriminative semantics, leading to further progress. Notably, even in data-limited situations, our method consistently demonstrates top performance. At the pixel level, with only 2 shots per class used for training, our method achieves improved average zero-shot performance compared to previous methods. With the full dataset, we set a new pixel-level SOTA with an AUROC of 93.4%. At the image level, our method is competitive with just 2 shots for training and establishes a new SOTA of 83.1% with 64 shots per class.

Unlike previous methods, our approach does not rely heavily on data resources to achieve top-tier performance. Comparison under different levels of data available, as shown in Fig. 5, reveals that our approach consistently outperforms other methods in general. Even with limited data, our model reaches competitive results, while other methods

Domain	Dataset	CLIP*	WinCLIP*	VAND*	MVFA-AD	AnomalyCLIP*	AdaCLIP	Ours			
		OpenCLIP	CVPR 2023	CVPRw 2023	CVPR 2024	ICLR2024	ECCV2024	-			
		Available training shots	-	-	full	full	full	2	16	64	full
Industrial	BTAD	30.6	32.8	91.1	90.1	93.3	90.8	92.8	94.4	96.5	97.0
	MPDD	62.1	95.2	94.9	94.5	96.2	96.6	96.3	96.5	96.3	96.7
	MVTec-AD	38.4	85.1	87.6	84.9	91.1	89.9	91.0	91.2	91.6	91.9
	VisA	46.6	79.6	94.2	93.4	95.4	95.5	93.4	93.8	94.0	95.5
Medical	Brain MRI	68.3	86.0	94.5	95.6	96.2	93.9	96.3	96.4	96.5	95.5
	Liver CT	90.5	96.2	95.6	96.8	93.9	94.5	97.3	97.7	97.7	97.8
	Retina OCT	21.3	80.6	88.5	90.9	92.6	88.5	94.2	95.1	94.4	95.5
	ColonDB	49.5	51.2	78.2	78.4	82.9	80.0	83.9	83.5	84.7	84.0
	ClinicDB	47.5	70.3	85.1	83.9	85.0	85.9	89.2	87.6	87.8	89.9
	Kvasir	44.6	69.7	80.3	81.9	81.9	86.4	82.1	84.6	85.2	87.2
	CVC-300	49.9	-	92.8	82.6	95.4	92.9	96.0	97.4	96.0	96.4
Average		49.9	74.7	89.3	88.5	91.3	90.4	92.0	92.6	92.8	93.4

Table 1. Pixel-level AUROC of zero-shot AD methods in Industrial and Medical domains. Method sources and the number of shots used for training are noted. Results of methods with * are copied from the papers or inferred from official weight. Best results are highlighted as **first**, **second** and **third**.

Domain	Dataset	CLIP&VAND*	WinCLIP*	MVFA-AD	AnomalyCLIP*	AdaCLIP	Ours				
		OpenCLIP	CVPR 2023	CVPR 2024	ICLR2024	ECCV2024	-				
		Available training shots	-	-	full	full	full	2	16	64	full
Industrial	BTAD	73.6	68.2	94.3	85.3	90.9	88.0	90.9	94.7	94.8	
	MPDD	73.0	63.6	70.9	73.7	72.1	63.6	78.3	75.7	75.1	
	MVTec-AD	86.1	91.8	86.6	90.9	90.0	85.9	89.7	92.0	90.5	
	VisA	66.4	78.0	76.5	82.1	84.3	78.4	84.0	84.1	84.6	
Medical	Brain MRI	58.8	66.5	70.9	83.3	80.2	84.3	80.4	83.4	80.2	
	Liver CT	54.7	64.2	63.0	61.6	64.2	69.4	68.1	69.2	69.7	
	Retina OCT	65.6	42.5	77.3	75.7	82.7	77.4	81.0	82.9	82.7	
	Average	68.3	67.8	77.1	78.4	80.6	78.1	81.8	83.1	82.5	

Table 2. Image-level AUROC of zero-shot AD methods in Industrial and Medical domains. Method sources and the number of shots used for training are noted. Results of methods with * are copied from the papers or inferred from official weight. Best results are highlighted as **first**, **second** and **third**.

display signs of underfitting. As data increases, our method maintains its lead, establishing a new SOTA at both pixel and image levels.

4.3. Visualization

To illustrate the alignment intuitively, we present visualization examples in Fig. 6 with original configuration for previous works. Although previous methods with can detect anomalous regions, our AA-CLIP demonstrates fewer false-negative predictions in both industrial and medical domains, accurately highlighting the correct anomaly regions.

4.4. Ablations Analysis

We conduct thorough ablation experiments of our refinement of both visual and text space, as shown in Tab. 3 and Fig. 7. The second row in Tab. 3, which mirrors the structure of VAND [6], serves as our baseline.

Image Space: As shown in Tab. 3 line “2.”, inserting the vallina linear adapter into transformer layers results in a significant decline in zero-shot performance, indicating the

Method	Avg. AUROC	
	Pixel-Level	Image-Level
CLIP	50.3	69.3
1. + Linear Proj. (VAND [6])	88.9	69.3
2. + Adapter	48.9(-40.0)	53.4(-15.9)
3. + Residual Adapter	91.3(+2.4)	80.7(+11.4)
4. + Residual Adapter	92.1(+3.2)	82.6(+13.3)
5. + Disentangle Loss	92.7(+3.8)	83.3(+14.0)

Table 3. **Ablation Study of Our Training Strategy with VisA-Trained 64-Shot Setup.** Our contributions are **bold**. While VAND uses linear projectors to improve AD performance, incorporating Residual Adapters further refines patch feature adaptation. Moreover, integrating our Disentangle Loss yields the best overall results.

damage of the original generalization ability of CLIP. Incorporating our Residual Adapters mitigates this issue (shown in line “3.”), enhancing performance while preserving original information stored in CLIP.

Text Space: The last two rows in Tab. 3 highlight the impact of our approach in equipping CLIP’s encoder with

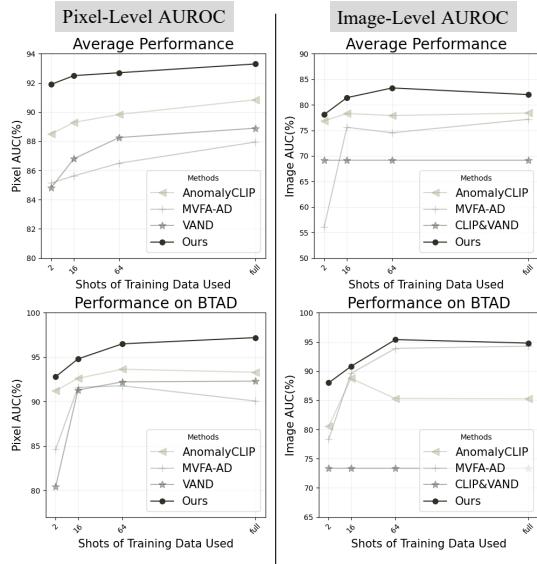


Figure 5. Average Results (Top) and Results on BTAD (Bottom) of Different methods Trained on 2-, 16-, 64-shot per Class and Full Data of VisA. Our method shows high fitting efficiency, achieving strong results across all data scales.

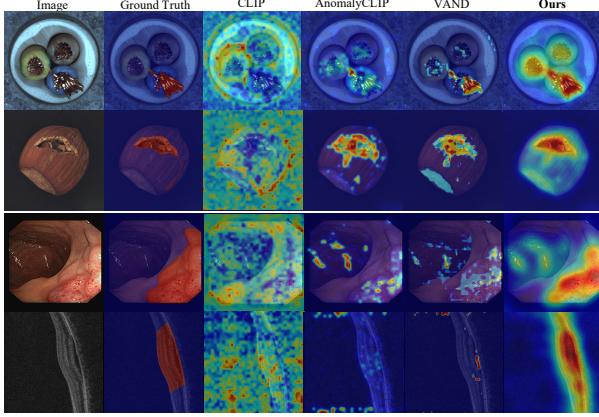


Figure 6. Visualization of Anomaly Localization Results of Original CLIP [41], AnomalyCLIP [57], VAND [6] and our AA-CLIP. Compared to previous methods, AA-CLIP demonstrates more reliable prediction capabilities in localizing anomaly.

anomaly-aware semantics. Line “4.” validates that, with AA-CLIP, the model’s ability to discriminate anomalies further improves, as the AA-CLIP’s text encoder provides a more precise semantic foundation. Adding Disentangle Loss leads to an additional improvement (shown in Line “5.”), especially at image-level, validating the necessity of independence between normal and anomaly anchors. These results underscore the crucial role of text space refinement in improved anomaly localization and classification.

Two-Stage Training: To validate the necessity of two-stage training, we adapt both text and image encoders together within one stage (also adopted by AdaCLIP). As shown

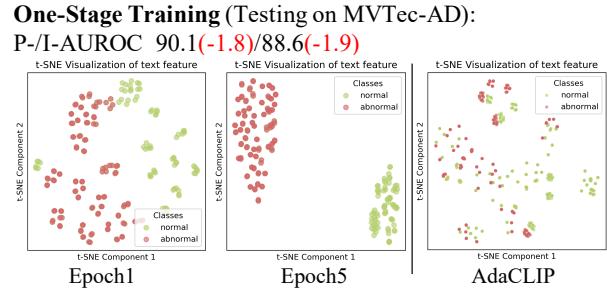


Figure 7. Visualization of Text Space from One-Stage Training and from AdaCLIP. During one-stage training, class information collapses easily, leading to damaged zero-shot performance.

in Fig. 7, one-stage model can easily exaggerate anomaly semantics and forget class information embedded in CLIP, damaging the model’s generalization ability. The two-stage training strategy allows controlled adaptation, preserving CLIP’s class-relevant knowledge in one end while adapting the other, as shown in Fig. 3.

5. Conclusion and Discussion

To our knowledge, this is the first work to explicitly analyze the intrinsic Anomaly Unawareness problem in CLIP. To tackle this issue, we propose a simple yet effective two-stage training strategy to embed anomaly-aware information into CLIP, enabling clear disentanglement of anomaly representations across both seen and novel classes. By leveraging residual adapters, our method preserves CLIP’s strong generalization ability, achieving outstanding zero-shot performance across multiple datasets.

Our adapted AA-CLIP, developed through this two-stage adaptation strategy, reveals the potential of refining CLIP’s feature space for improved performance in downstream applications. Beyond addressing anomaly unawareness, our work also provides a potential foundation for tackling other “unawareness” issues within CLIP. These may include limitations in context-awareness or specificity to domain-relevant nuances, suggesting further applications of our method in expanding CLIP’s adaptability across diverse tasks. Additionally, we observe signs of overfitting with full-shot training, suggesting potential saturation during CLIP adaptation and warranting further investigation.

Acknowledgement

This work is supported by Natural Science Foundation of China under Grant 62271465, Suzhou Basic Research Program under Grant SYG202338, Open Fund Project of Guangdong Academy of Medical Sciences, China (No. YKY-KF202206), and Jiangsu Province Science Foundation for Youths (NO. BK20240464).

References

- [1] Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection, 2024. 6
- [2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtac ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 1, 3, 6
- [3] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012. 6
- [4] 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*, 43:99–111, 2015. 6
- [5] 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 *European Conference on Computer Vision*, pages 55–72. Springer, 2025. 2, 3, 5, 6
- [6] 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. *arXiv preprint arXiv:2305.17382*, 2023. 2, 3, 5, 6, 7, 8
- [7] Xuhai Chen, Jiangning Zhang, Guanzhong Tian, Haoyang He, Wuhan Zhang, Yabiao Wang, Chengjie Wang, and Yong Liu. Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection. *arXiv preprint arXiv:2311.00453*, 2023. 3
- [8] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dermoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115*, 2022. 2
- [9] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 1, 2, 6
- [10] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9737–9746, 2022. 1, 2, 6
- [11] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 2
- [12] Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection—a survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37, 2021. 1
- [13] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. 3
- [14] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 98–107, 2022. 1, 2
- [15] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. Diad: A diffusion-based framework for multi-class anomaly detection. *arXiv preprint arXiv:2312.06607*, 2023. 2
- [16] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11375–11385, 2024. 2
- [17] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11375–11385, 2024. 3, 5, 6
- [18] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023. 3, 6
- [19] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*, pages 66–71. IEEE, 2021. 6
- [20] Ruixiang Jiang, Lingbo Liu, and Changwen Chen. Clip-count: Towards text-guided zero-shot object counting. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4535–4545, 2023. 2
- [21] Zeeshan Khan, Makarand Tapaswi, et al. Figclip: Fine-grained clip adaptation via densely annotated videos. *arXiv preprint arXiv:2401.07669*, 2024. 2
- [22] Daehyun Kim, Sungyong Baik, and Tae Hyun Kim. Sanflow: Semantic-aware normalizing flow for anomaly detection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2
- [23] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2
- [24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.
- [25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with

- frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2
- [26] Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [27] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 5
- [28] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15305–15314, 2023. 2, 3
- [29] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023. 2, 3
- [30] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411, 2023. 1, 2
- [31] Ruiying Lu, YuJie Wu, Long Tian, Dongsheng Wang, Bo Chen, Xiyang Liu, and Ruimin Hu. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2310.14228*, 2023. 2
- [32] Wenxin Ma, Qingsong Yao, Xiang Zhang, Zhelong Huang, Zihang Jiang, and S.Kevin Zhou. Towards accurate unified anomaly segmentation. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 1342–1352, 2025. 2
- [33] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE, 2021. 6
- [34] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2
- [35] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15579–15591, 2023. 2
- [36] Amin Karimi Monsefi, Kishore Prakash Sailaja, Ali Alilooee, Ser-Nam Lim, and Rajiv Ramnath. Detailclip: Detail-oriented clip for fine-grained tasks. *arXiv preprint arXiv:2409.06809*, 2024. 3
- [37] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3170–3180, 2023. 2
- [38] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094, 2021. 2
- [39] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Grutowicz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 164–169, New York, NY, USA, 2017. ACM. 6
- [40] 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. *arXiv preprint arXiv:2407.12276*, 2024. 2, 3, 6
- [41] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 8
- [42] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022. 2, 6
- [43] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017. 2
- [44] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022. 2
- [45] Yingtian Tang, Yutaro Yamada, Yoyo Zhang, and İlker Yıldırım. When are lemons purple? the concept association bias of vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14333–14348, 2023. 2
- [46] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzał, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017(1):4037190, 2017. 6
- [47] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for anomaly detection. *arXiv preprint arXiv:2103.04257*, 2021. 2
- [48] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 2, 3
- [49] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic

- segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 3
- [50] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18938–18949, 2023. 2
- [51] Qingsong Yao, Li Xiao, Peihang Liu, and S Kevin Zhou. Label-free segmentation of covid-19 lesions in lung ct. *IEEE transactions on medical imaging*, 40(10):2808–2819, 2021. 2
- [52] Xincheng Yao, Chongyang Zhang, Ruqi Li, Jun Sun, and Zhenyu Liu. One-for-all: Proposal masked cross-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4792–4800, 2023. 2
- [53] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022. 2, 6
- [54] Zhiyuan You, Kai Yang, Wenhan Luo, Lei Cui, Yu Zheng, and Xinyi Le. Adtr: Anomaly detection transformer with feature reconstruction. In *International Conference on Neural Information Processing*, pages 298–310. Springer, 2022. 2
- [55] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kun-chang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022. 3
- [56] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3914–3923, 2023. 1, 2
- [57] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023. 3, 5, 6, 8
- [58] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023. 2, 3
- [59] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018. 2
- [60] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022. 6