This CVPR Findings paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

DA-CLIP: Mitigating Granularity Mismatch in Zero-Shot Anomaly Detection via Decoupled Text-Visual Alignment

Jianqin Liu[1] Peng Wang[3*] Junming Huang[1] Xue Zhou[2,1] Li Yu[1]
[1]School of Automation Engineering, University of Electronic Science and Technology of China
[2]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China
[3]School of Computer Science and Engineering, University of Electronic Science and Technology of China
liujianqin1@gmail.com p.wang6@hotmail.com

Abstract

Recently, Vision-Language Models (VLMs) like CLIP have been widely adopted for zero-shot anomaly detection (ZSAD). Existing methods commonly perform a joint alignment of a single text feature with both global and local visual features. However, they overlook a core challenge: the granularity mismatch between the image-level classification and pixel-level segmentation tasks. In this work, we present our observation that joint alignment leads to two unexpected phenomena: 1) Performance degrades and is even inferior to the original CLIP, despite a clear separation between normal and abnormal text features. 2) The text features show a preference for aligning with local visual features, which enlarges the modality gap with global features and consequently harms classification performance. To address these issues, we propose a simple yet effective Decoupled Alignment framework (DA-CLIP). Our framework decouples feature alignment into two separate spaces using two independent text encoders, preventing mutual interference and effectively mitigating the granularity mismatch. Furthermore, to enhance the focus of text features on anomaly semantics rather than foreground object semantics, while preserving discriminability, we introduce a Decoupled Asymmetric Refinement (DAR) block, which dynamically refines the text features by incorporating fine-grained image priors. Extensive experiments on 11 industrial and medical datasets demonstrate that our method achieves state-of-the-art performance.

1. Introduction

Recently, Contrastive Language-Image Pretraining (CLIP) model [30] has demonstrated remarkable generalization capabilities in zero-shot learning [20, 34], driving widespread adoption of zero-shot anomaly detection (ZSAD) in domains such as industrial defect inspection [6, 25, 29, 41] and medical diagnosis [17, 39]. ZSAD aims to train exclusively on auxiliary datasets and then detect previously unseen anomalies in the target domain. Unlike traditional AD methods, ZSAD does not require a large corpus of normal target domain samples for prior training, thus overcoming the inability of traditional methods to detect novel or rare defect classes. Consequently, ZSAD represents a more meaningful and inherently challenging task.

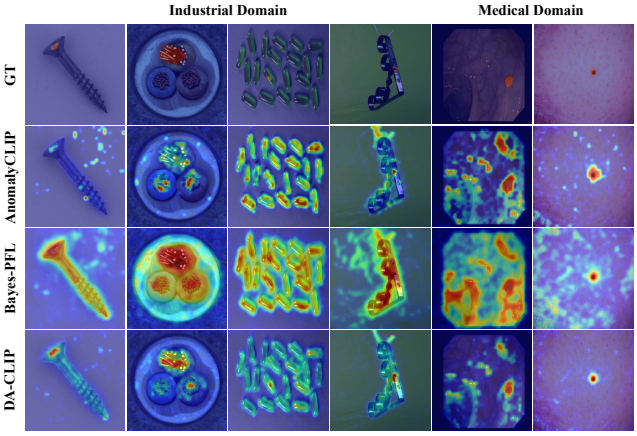


Figure 1. **Comparison with Existing State-of-the-Art Methods.** Ground truth regions are highlighted in red in the first row. DA-CLIP delivers more accurate segmentation results, reducing both false negatives and false positives.

Recent efforts [7, 8, 18, 25, 28, 29, 41] have explored adapting CLIP's strong generalization to AD domains. Initial researches primarily focus on how to generalize CLIP's pre-trained knowledge to AD tasks. Some methods [6, 11, 25, 28, 41] improve ZSAD performance by using adapters or inserting learnable vectors to achieve a tight alignment between visual and text features. Subsequent method [25] explores the limitations of CLIP's text features regarding their lack of anomaly awareness. However,

*Corresponding author

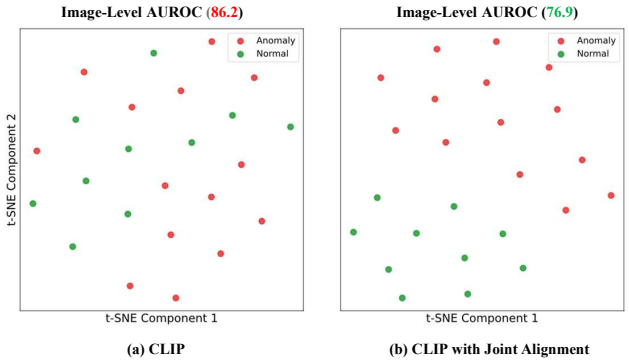


Figure 2. **t-SNE Visualization of Text Features.** While the original CLIP's normal and abnormal text features are entangled, joint alignment effectively separates them. However, the classification performance paradoxically declines.

these approaches overlook a critical issue: AD tasks are composed of both a coarse-grained classification task and a fine-grained segmentation task. These methods jointly align the adapted text features with both global image features and patch features to optimize classification and segmentation tasks simultaneously. Based on this fundamental understanding, we observe that aligning an already anomaly-aware text feature with global image features paradoxically degrades classification performance, making it even inferior to the original CLIP, as shown in Fig. 2.

What hinders CLIP's performance in ZSAD? To assess the limitations of CLIP in ZSAD, we analyze its embedding space, as illustrated in Fig. 3. Specifically, we observe in Fig. 3(a) that within the original CLIP's embedding space, the normal and abnormal text features are entangled. This, combined with a significant modality gap from the visual features, leads to poor performance in both classification and segmentation (Fig. 3(d)). While the joint alignment approach establishes a clear distinction between normal and abnormal semantics in the text space, the text features show a strong preference for aligning with patch features. This alignment inadvertently widens the modality gap with global features, causing a degradation in classification performance, as depicted in Fig. 3(b).

This critical observation reveals the fundamental reason for CLIP struggles in ZSAD: the joint alignment paradigm disregards the inherent granularity mismatch between the classification and segmentation sub-tasks by compelling their optimization within a shared embedding space. Therefore, the question arises: *is it possible to resolve this conflict by decoupling the alignment within a unified framework, using separate adapters for a non-interfering, simultaneous optimization of both tasks?*

Our solution. Specifically, we process the text prompts through two separate text encoders to obtain task-specific text features. One feature is generated by a trainable text encoder and is exclusively aligned with global image features, focusing on the classification task. Conversely, the other feature, generated by a frozen text encoder, is aligned with patch features via a learnable linear projection layer to guide it towards anomalous regions. This decoupled alignment framework ensures that each text feature achieves precise alignment with its corresponding visual feature within a dedicated embedding space. As a result, it prevents the phenomenon of preferential alignment and narrows the gap between modalities, as depicted in Fig. 3(c).

Furthermore, as evidenced by the anomaly maps in Fig. 3(f), we observe significant false positive regions. We attribute this primarily to the inherent limitation of the frozen text features, which tends to focus on semantic modeling of foreground objects rather than identifying anomalies. To overcome this limitation, we introduce a Decoupled Asymmetric Refinement (DAR) block, which leverages the decoupled text embeddings via an additional independent text encoder and dynamically refines them by incorporating fine-grained visual features. This targeted refinement further unlocks the segmentation potential of ZSAD, yielding significantly more precise anomaly localization (Fig. 3(g)). In summary, our contributions are as follows:

- We identify a critical granularity mismatch problem in ZSAD, revealing the underlying reason for CLIP's limited performance.
- To address this issue, we propose DA-CLIP, a simple yet effective decoupled alignment framework. It achieves specialized alignment by decoupling the task optimization process into two embedding spaces.
- We introduce a DAR block that refines text embeddings using fine-grained visual features, further boosting segmentation performance.
- Extensive empirical results demonstrate that DA-CLIP achieves state-of-the-art (SOTA) performance in 11 industrial and medical benchmark datasets.

2. Related Work

2.1. Traditional Anomaly Detection

Most existing anomaly detection methods rely on an unsupervised paradigm, training exclusively on large collections of normal samples. Reconstruction-based methods [14–16, 35, 36, 38] reconstruct test samples and detect anomalies by measuring reconstruction error. Knowledge distillation-based methods [10, 23, 33] identify anomalies through the feature discrepancy between teacher and student networks. Synthesis-based methods [21, 32, 37, 40] generate synthetic anomalous examples during training to enhance the model's sensitivity to defects. Although they often achieve strong results, their success critically depends on the quantity and diversity of normal examples. Consequently, performance degrades substantially when con-

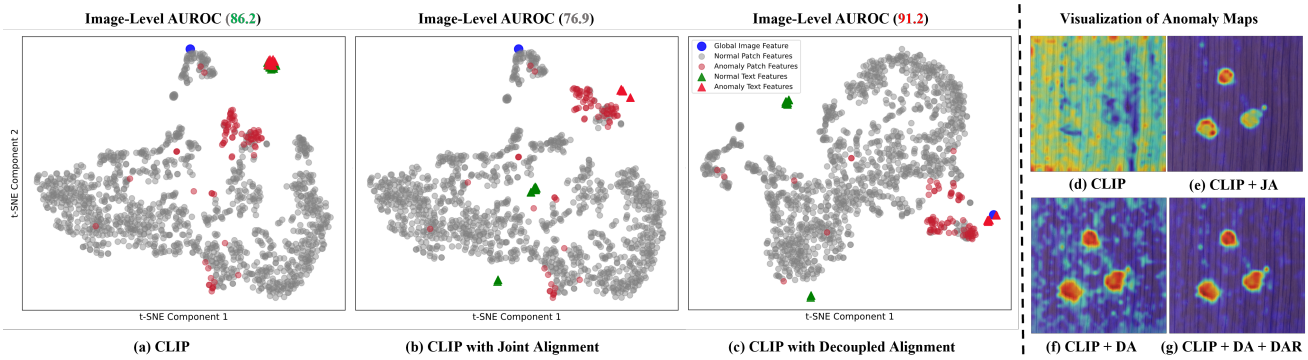


Figure 3. **t-SNE Visualization of CLIP Embedding Space and Anomaly Maps.** (a) In the original CLIP, text features are entangled and misaligned with visual features. (b) Joint alignment (JA) causes text features to preferentially align with patch features, degrading classification. (c) Our decoupled alignment (DA) effectively aligns each text feature with its corresponding visual feature, thereby bridging the modality gap. (d-g) Anomaly maps visualize the progressive impact of each component.

fronted with novel defect types or rare diseases. Moreover, training a separate model for each category incurs prohibitive computational and time costs.

2.2. Zero-shot Anomaly Detection

Due to CLIP's powerful generalization, ZSAD has garnered significant attention. WinCLIP [18] is the first to align composite textual prompts, with aggregated multilayer visual features. APRIL-GAN [7] and CLIP-AD [8] employ a learnable adapter to project visual features into the joint embedding space, enabling targeted alignment between textual and visual features and thereby enhancing segmentation performance. However, the aforementioned methods neglect to explicitly optimize the textual space, resulting in entangled embeddings of normal and anomalous semantics, which limit CLIP's classification performance. To address this issue, AnomalyCLIP [41] and FiLo [12] introduce learnable vectors to adapt the textual space for better anomaly discrimination. AdaCLIP [6] proposes a hybrid of static and dynamic prompts to jointly optimize both the textual and visual spaces. AA-CLIP [25] is the first to reveal CLIP's inherent lack of anomaly awareness and inserts residual adapters into both the text and image encoders, enabling clear separation of normal and anomalous semantics in both spaces. Bayes-PFL [29] presents a Bayesian prompt flow learning strategy to regularize both context and state text spaces, thereby improving generalization to unseen classes. However, these methods overlook a critical issue: AD tasks are composed of a coarse-grained classification task and a fine-grained segmentation task, which have an inherent granularity mismatch. Therefore, directly performing a joint alignment of text and visual features limits their performance.

Unlike previous methods, we are the first to identify the true reason for CLIP's limitations in the ZSAD task. We propose a simple yet effective decoupled alignment framework that decouples the alignment process into two independent subspaces. This architecture compels each of the two text features to learn task-specific semantics independently, effectively preventing the optimization conflict.

3. Method

3.1. Problem Formulation

ZSAD aims to train a generalizable anomaly detection model capable of operating effectively on entirely unseen domains and anomaly types, without requiring any target domain samples or prior knowledge. Following the recent setting of ZSAD tasks, the model is trained exclusively on the industrial dataset $\mathcal{D}_{train}$ and evaluated directly on unseen categories from different test datasets $\{\mathcal{D}^1_{test}, \mathcal{D}^2_{test}, ..., \mathcal{D}^n_{test}\}$, where the categories and defect types in each $\mathcal{D}^i_{test}$ are distinct from $\mathcal{D}_{train}$. The evaluation comprises an image-level classification task to determine normal/anomalous samples and a pixel-level segmentation task to localize anomalous regions.

3.2. Key Observations

Although prior methods have made some progress in ZSAD, they commonly overlook the inherent granularity mismatch problem of the AD tasks, and the joint alignment strategy exhibits certain limitations. To reveal this issue, we conduct visualization analysis of the embedding space. We found that while the joint alignment approach successfully creates a clear separation between normal and abnormal text features, the image-level AUROC metric drops significantly from 86.2% to 76.9%. This phenomenon reveals the problem: the text feature preferentially aligns with the more numerous local features, thereby neglecting the global image feature. This preference leads to a further widening of the modality gap between the text and global features (Fig. 3(b)).

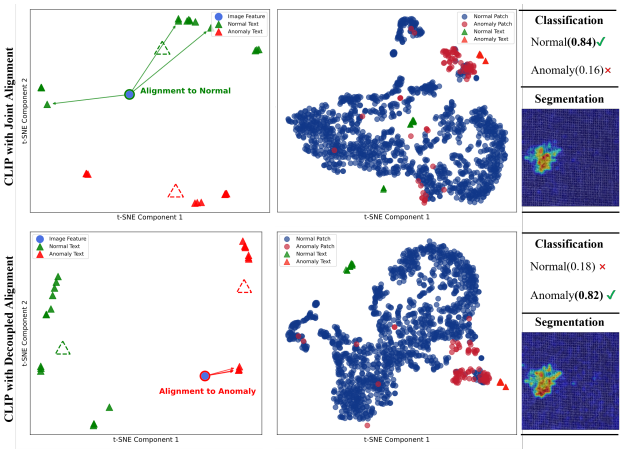


Figure 4. **t-SNE Visualizations of Text-Visual Alignment.** The dashed triangle represents the centroid of the text features. Although joint alignment can effectively locate anomalies within an image, it fails to determine if the image itself is anomalous. In contrast, decoupled alignment succeeds in both localizing and identifying anomalies.

To further analyze how this preferential alignment leads to failure, Fig. 4 provides a visual analysis within the two visual spaces. Under joint alignment (Fig. 4(top)), although the model locates the anomalous regions in the image, the global feature is pushed further away from the abnormal text feature in the global space. Instead, it moves closer to the normal text feature, causing the entire image to be misclassified as normal. This alignment paradigm forces a single text feature to reconcile two tasks of fundamentally different granularities, a practice that is fundamentally detrimental to ZSAD.

3.3. Overview

Fig. 5 illustrates the overall framework of DA-CLIP, which consists of two key components: 1) the decoupled alignment framework decouples text-visual alignment process to independently optimize classification and segmentation tasks, 2) the DAR branch directs text features to focus on anomalous regions rather than foreground object semantics. Specifically, given an input image $x \in \mathbb{R}^{h\times w\times 3}$, the image encoder produces the global image features $V_{cls} \in \mathbb{R}^C$ and the patch features $V_p^i \in \mathbb{R}^{H\times W\times C}, i \in \{1,2,3,4\}$, where $H = h/patchsize, W = w/patchsize$ and C represents the embedding dimension. V_p^i are passed through a single layer to produce F_p^i. Concurrently, text prompts with standard ZSAD protocols are input into both the text encoder integrated with token adapter and the original CLIP text encoder, yielding two text features, T_c and $T_s \in \mathbb{R}^{2\times C}$. T_c are aligned with V_{cls} to compute anomaly scores, T_s are aligned with F_p^i to generate a coarse anomaly map M_1. Furthermore, we introduce the DAR block to adjust text features for precise segmentation. Note that, to avoid impairing the original classification capability, we additionally decouple a dedicated text embedding T_m. Through the refinement module, we dynamically update T_m and align it with V_p^i to generate a fine anomaly map M_2. Finally, we jointly optimize the coarse map M_1 and the refined map M_2 to supervise the segmentation task.

3.4. Decoupled Alignment Framework

Decoupled Alignment. Motivated by Sec. 3.2, we propose a novel decoupled alignment framework to synergistically unlock the potential of ZSAD performance. Specifically, we design object-agnostic handcrafted text prompts (eg. "a photo of normal [object]" and "a photo of damaged [object]"), which are fed into two text encoders Ψ_1 and Ψ_2 respectively, generating decoupled text features.

$$\begin{aligned} T_c &= [T_c^a, T_c^n] = \Psi_1([t_a, t_n]), \\ T_s &= [T_s^a, T_s^n] = \Psi_2([t_a, t_n]), \end{aligned} \tag{1}$$

where T^a and $T^n \in \mathbb{R}^C$ denote the anomalous and normal text features. t_a and t_n represent the anomalous and normal text prompt. $[\cdot,\cdot]$ denotes concatenation along rows.

Unlike existing joint alignment paradigms, we assign decoupled text features to classification and segmentation independently, which prevents confusion caused by their granularity mismatch and avoids any related performance drop. Specifically, for the classification task, we measure the cosine similarity between T_c and the global image feature V_{cls} to compute the image-level anomaly score, as defined in Eq. (2):

$$S_{cls} = \mathrm{Cos}\, Sim(V_{cls}, T_c), \tag{2}$$

where $S_{cls} \in \mathbb{R}^2$ denotes the classification prediction.

For the segmentation task, we first project V_p^i into a unified embedding space using a learnable linear layer to produce F_p^i, thereby facilitating better alignment between text features and patch features.

$$F_p^i = Linear_i(V_p^i), i \in \{1,2,3,4\} \tag{3}$$

However, directly computing the cosine similarity between T_c and F_p^i can cause rapid overfitting on the auxiliary dataset, which undermines CLIP original generalization and degrades performance [11, 28]. Accordingly, we instead compute the similarity between the additional text features T_s and F_p^i.

$$M_1 = \sum_{\mathrm{i}=1}^{4} \mathrm{Cos}\, Sim(F_p^i, T_s), \tag{4}$$

where $M_1 \in \mathbb{R}^{HW\times 2}$ denotes the segmentation prediction. Ultimately, our simple yet effective decoupled alignment framework successfully achieves a targeted alignment between text and visual features, as shown in Fig. 4 (bottom).

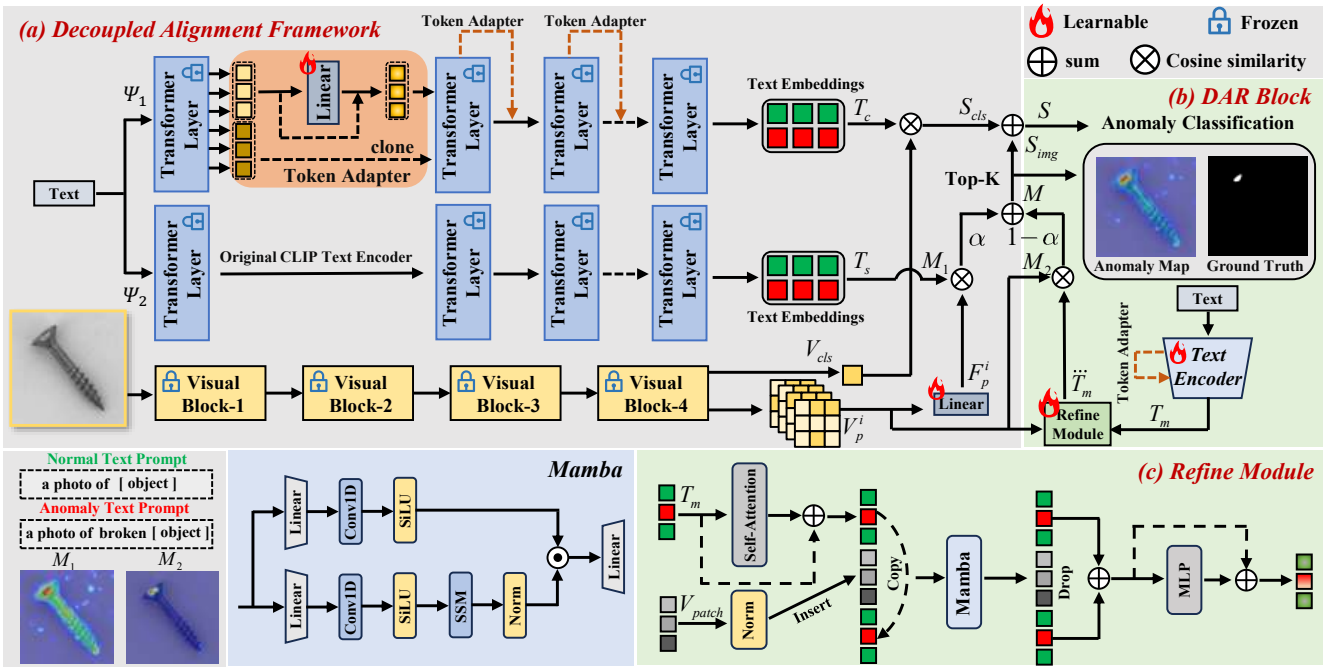


Figure 5. **Overview of DA-CLIP.** We introduce a decoupled alignment framework that solves the granularity mismatch problem between classification and segmentation by decoupling the alignment process into two separate subspaces (Sec. 3.4). We then introduce the DAR block to produce fine-grained text features and generate anomaly maps, effectively reducing false positives (Sec. 3.5).

Token Adapter. To better generalize CLIP's text semantics to the anomaly semantic space, previous works [6, 12, 28, 41] have attempted to refine the text space by adding extra learnable token embeddings into its text encoder. However, this approach often overfits to the training data, damaging CLIP's original pre-trained knowledge and thus leading to poor generalization on new datasets. Additionally, the number of learnable tokens and their insertion layers must be set empirically. To address these problems, we propose a lightweight token adapter and insert it into all layers of the text encoder.

The output token embedding $x^i \in \mathbb{R}^{M\times C}$ of CLIP's i-th tokenizer and the token embedding layer is fed into the i-th adapter, outputting adapter token embedding $x^i_{adapted}$, as shown in Eq. (5).

$$x^i_{adapted} = x^i + Act(Linear(x^i)), \tag{5}$$

where $Act(\cdot)$ is an activation function. The residual connection preserves the original semantics within the textual space, which facilitates generalization to unseen classes.

3.5. Decoupled Asymmetric Refinement Block

Although our decoupled alignment framework achieves strong segmentation performance, we observe persistent false positives on the anomaly maps (Fig. 3(f)). We attribute this to CLIP's training process: the non-fine-tuned text features that focus on foreground object semantics rather than anomalous regions, which cause misidentification of objects as anomalies during alignment. To address this issue, we introduce Decoupled Asymmetric Refinement (DAR) block as shown in Fig. 5(b). In order to preserve the original anomaly awareness in text spaces, we introduce an additional text encoder integrated with token adapters, which focuses specifically on aligning text features with patch features rather than with the global image features.

Its core innovation is a refinement module that dynamically adjusts text features using fine-grained visual features, establishing a robust text-visual alignment. The refinement module uses a hybrid architecture which combines the strengths of self-attention and Mamba: the attention block captures semantic discrepancies the normal and abnormal text features, while the Mamba block dynamically updates the text features based on visual features, resulting in fine-grained text features.

As shown in Fig. 5(c), we first employ a self-attention mechanism to explore the relationship between normal and anomalous text features.

$$\dot{T}_m = T_m + SA(T_m, T_m, T_m), \tag{6}$$

where SA (Q, K, V) denotes self-attention, the order of input is query, key, value.

Although Mamba excels at modeling long sequences, its use in cross-modal tasks remains unexplored. To harness Mamba's unidirectional scanning order, we concatenate text

features to both ends of the patch tokens, thereby maximizing the incorporation of fine-grained visual features.

$$[\Delta\dot{T}_m, \Delta V_p^i, \Delta\dot{T}_m^{copy}] = \text{Mamba}([\dot{T}_m, V_p^i, \dot{T}_m^{copy}]), \quad (7)$$

$$\ddot{T}_m = \dot{T}_m + \Delta\dot{T}_m + \Delta\dot{T}_m^{copy}, \quad (8)$$

$$\dddot{T}_m = \ddot{T}_m + \text{MLP}(\ddot{T}_m), \quad (9)$$

We adopt Mamba from [13] in our design, MLP denotes a multilayer perceptron. Subsequently, we compute the similarity (same as Eq. (4)) between these refined text features $\dddot{T}_m$ and V_p^i to produce precise anomaly map M_2. Importantly, we employ an asymmetric design by using V_p^i as the patch tokens instead of F_p^i, jointly optimizing text and visual spaces causes rapid overfitting on the training dataset, which impairs its generalization ability on unseen classes, as validated in Tab. 3.

3.6. Training and Inference

Loss Function. We follow previous work [12, 25, 41] in jointly optimizing both global and local alignment objectives. Consequently, the total loss function of DA-CLIP is calculated as:

$$\begin{aligned}\mathcal{L}_{total} = &\underbrace{\mathcal{L}_{CE}(S_{cls}, y) + \mathcal{L}_{Focal}(M_1, G) + \mathcal{L}_{Dice}(M_1, G)}_{\text{Decoupled Alignment Framework}} \\ &+ \underbrace{\mathcal{L}_{Focal}(M_2, G) + \mathcal{L}_{Dice}(M_2, G)}_{\text{DAR Block}},\end{aligned} \quad (10)$$

where $\mathcal{L}_{CE}$ denotes cross entropy loss, $\mathcal{L}_{Focal}$ denotes focal loss [22], $\mathcal{L}_{Dice}$ denotes dice loss [26], y denotes true label, and G represents ground truth. M_1 and M_2 are reshaped and upsampled to align with segmentation mask G.

Anomaly Map and Anomaly Score. The final anomaly map is generated by computing a weighted sum of M_1 and M_2, $M = (1-\alpha)M_1 + \alpha M_2$, where α is a hyperparameter to balance the importance of different anomaly maps. For image-level scoring, we incorporate pixel-level score to provide more robust guidance. Unlike previous approaches that rely solely on the single maximum response of the anomaly map, which is prone to noise and false positives. We instead extract the top-k responses from the anomaly map to represent the pixel-level anomaly score S_{img}. Next we combine S_{img} with the classification prediction S_{cls}. Formally, the final image-level anomaly score S is defined as:

$$\begin{aligned}S &= S_{cls} + S_{img} \\ &= S_{cls} + \text{Mean}(\text{Top-}k(M)).\end{aligned} \quad (11)$$

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our method on 11 authoritative datasets across industrial and medical domains. Specifically, for industrial anomaly detection, we use MVTec AD [4], VisA [42], BTAD [27], MPDD [19], and DTD-Synthetic [3]. For the medical domain, we include brain tumor detection datasets HeadCT [31], BrainMRI [31] and Br35H [1], polyp segmentation dataset CVC-ClinicDB [5] with pixel-level labels, the skin cancer dataset ISIC [9], and lung cancer segmentation dataset MSD [2]. Following the standard ZSAD protocol, we fine-tune on the auxiliary dataset and directly evaluate on other datasets. Consistent with prior work [24, 41], we use MVTec AD as our auxiliary dataset. To evaluate performance on MVTec AD, we train the model on VisA dataset. We present the evaluation results obtained by using medical datasets as auxiliary data in the supplementary material.

Evaluation Metrics. To comprehensively assess model performance, we report image-level Area Under the Receiver Operating Characteristic (AUROC) and the average precision (AP) for anomaly classification, and pixel-level AUROC and Per-Region Overlap (PRO) for anomaly segmentation.

Implementation Details. In this paper, we use the CLIP model (`VIT-L/14@336px`) pretrained by OpenAI as our backbone. The image resolution is 518×518. We extract features from the 6th, 12th, 18th, and 24th layers of the image encoder as patch features. α is set to 0.7. We train the model for 25 epochs with a batch size of 16. The token adapters are optimized at a learning rate of 1e-5, and the other learnable parameters are updated at 5e-6. Parameters are updated by Adam optimizers. All experiments are conducted on a single NVIDIA GeForce RTX 3090. More details can be found in the supplementary material.

4.2. Comparison with SOTA Methods

We compare our method against five SOTA methods: WinCLIP [18], APRIL-GAN [7], AnomalyCLIP [41], TPS [24] and Bayes-PFL [29]. To ensure a fair comparison, all methods are evaluated under identical settings to our approach (the same backbone, input image resolution, dataset configurations). More details about SOTA methods can be found in the supplementary material.

Quantitative Comparisons. As shown in Tab. 1, DA-CLIP achieves state-of-the-art results across image-level and pixel-level metrics on diverse datasets. Our method demonstrates superior classification and segmentation performance compared to existing approaches. Specifically, for the classification task on BTAD, our method achieves relative gains of 3.3% in image-level AUROC and 6.1% in AP over the second-best approach. For segmentation, on the industrial anomaly benchmarks VisA and BTAD, our method outperforms the second-best approach by 2.1% and 3.9% in pixel-level PRO, respectively. APRIL-GAN implements text–visual alignment using only a simple linear projection, which fails to imbue the text embedding space with anomaly awareness and consequently results in poor classi-

Task	Domain	Dataset	WinCLIP [18]	APRIL-GAN [7]	AnomalyCLIP [41]	TPS [24]	Bayes-PFL [29]	**DA-CLIP**
Image-level (AUROC, AP)	Industrial	VisA	(78.1, 81.2)	(78.0, 81.4)	(82.1, 85.4)	(83.4, 85.7)	(87.0, 89.2)	(88.5, 90.4)
		MVTec AD	(91.8, 96.5)	(86.1, 93.5)	(91.5, 96.2)	(89.5, 95.1)	(92.2, 96.0)	(91.9, 96.2)
		BTAD	(68.2, 70.9)	(73.9, 70.1)	(88.3, 87.3)	(88.2, 90.0)	(90.5, 89.1)	(93.8, 95.2)
		DTD-Synthetic	(93.2, 92.6)	(85.7, 94)	(93.5, 97.0)	(93.4, 96.7)	(95.3, 98.2)	(97.0, 98.9)
		MPDD	(63.6, 69.9)	(73.0, 80.2)	(77.0, 82.2)	(73.4, 74.7)	(76.9, 78.7)	(76.0, 82.8)
	Medical	Br35H	(80.5, 82.2)	(93.1, 92.9)	(94.6, 94.7)	(96.2, 96.1)	(94.0, 93.0)	(96.6, 96.5)
		BrainMRI	(86.6, 91.5)	(89.3, 90.9)	(90.3, 92.2)	(92.4, 93.8)	(91.1, 91.4)	(93.2, 94.5)
		HeadCT	(81.8, 80.2)	(89.1, 89.4)	(92.5, 90.6)	(92.3, 91.8)	(88.3, 87.6)	(92.6, 91.9)
Pixel-level (AUROC, PRO)	Industrial	VisA	(79.6, 56.8)	(94.2, 86.8)	(95.4, 86.0)	(94.4, 89.3)	(95.5, 89.6)	(95.5, 91.7)
		MVTec AD	(85.1, 64.6)	(87.6, 44.0)	(90.9, 81.1)	(86.6, 74.8)	(91.8, 87.8)	(91.2, 87.3)
		BTAD	(72.7, 27.3)	(89.3, 68.7)	(94.0, 74.7)	(92.2, 76.6)	(92.3, 78.7)	(94.0, 82.6)
		DTD-Synthetic	(83.9, 57.8)	(95.2, 87.3)	(97.9, 92.4)	(97.9, 93.4)	(98.2, 94.6)	(98.6, 95.8)
		MPDD	(76.4, 48.9)	(94.3, 83.8)	(96.4, 88.9)	(97.2, 90.8)	(97.1, 91.3)	(97.3, 92.7)
	Medical	ISIC	(83.3, 55.1)	(89.4, 77.2)	(89.9, 79.9)	(89.8, 80.8)	(89.7, 84.8)	(90.6, 83.9)
		ClinicDB	(51.2, 13.8)	(80.5, 60.7)	(83.2, 68.3)	(86.0, 71.5)	(84.8, 66.9)	(85.5, 71.1)
		MSD	(88.5, 60.8)	(92.3, 73.1)	(91.3, 67)	(93.1, 73.2)	(92.4, 72.1)	(93.9, 76.4)

Table 1. Comparison with existing state-of-the-art methods. The best results are marked in red, while the second-best are indicated in blue.

fication performance. AnomalyCLIP, TPS, and Bayes-PFL focus on crafting complex text prompts. In contrast, DA-CLIP employs simple text prompts without complex designs, yet achieves ZSAD performance. Notably, when DA-CLIP is fine-tuned exclusively on industrial datasets, it still achieves remarkable performance on medical datasets, highlighting its strong cross-domain generalization capability. We report the number of learnable parameters and the inference time of DA-CLIP. As shown in Tab. 2, despite the introduction of two additional text encoders, DA-CLIP still achieves a competitive parameter count and inference speed while delivering state-of-the-art performance.

Qualitative Comparisons. To more intuitively demonstrate segmentation performance, Fig. 1 provides visualizations between DA-CLIP, AnomalyCLIP, and Bayes-PFL across different datasets. Our DA-CLIP delivers more precise and clearly defined anomaly maps. In contrast, AnomalyCLIP can fail to detect certain defects, while Bayes-PFL generates extensive false positives, particularly when dealing with objects such as screws, cables, and colon samples. More visualizations are in the supplementary material.

4.3. Further Analysis

Granularity Mismatch Problem. To validate whether this problem exists in other methods, we analyze AnomalyCLIP as shown in Fig. 6. We find that AnomalyCLIP's text features preferentially align with patch features, creating a large modality gap with global visual features, which limits its classification performance. In contrast, DA-CLIP achieves targeted alignment. More visual analyses are provided in the supplementary material.

Token Adapter vs. Learnable Token. We investigate the strategy of optimizing the text space using learnable tokens within AnomalyCLIP. To ensure a fair comparison, we replace these learnable tokens with Token Adapters, while maintaining identical configurations for all other components. Notably, we only use simple handcrafted text prompts. These modified versions are referred to as AnomalyCLIP+, which demonstrates improved performance across nearly all benchmark datasets, as quantified by image-level AUROC (see Fig. 7). This significant enhancement demonstrates that the token adapter is more effective than learnable tokens in adapting the text space.

Method	Image-level		Pixel-level		Learnable Param(M)	Time (ms)
	AUROC	AP	AUROC	PRO		
APRIL-GAN	78.0	81.4	94.2	86.8	**3.1**	**105.0**
AnomalyCLIP	82.1	85.4	95.4	86.0	5.6	153.0
TPS	83.4	85.7	94.4	89.3	29.1	125.2
Bayes-PFL	87.0	89.2	95.5	89.6	27.4	365.9
DA-CLIP	**88.5**	**90.4**	**95.5**	**91.7**	25.4	124.6

Table 2. Comparison of learnable param and average inference time on VisA. The best results are shown in **bold**.

Method	VisA				MVTec AD			
	Pixel		Image		Pixel		Image	
	AUROC	PRO	AUROC	AP	AUROC	PRO	AUROC	AP
CLIP	36.9	7.3	78.2	81.3	25.6	3.1	86.2	94.5
+ Joint Alignment	95.2	90.3	69.3	73.9	89.3	83.4	76.9	88.6
+ Decoupled Alignment	95.4	91.2	87.5	89.5	90.9	86.8	91.2	95.8
+ DAR Block	**95.5**	**91.7**	**88.5**	**90.4**	**91.2**	**87.3**	**91.9**	**96.2**

Table 3. Ablation study on Our Key Design Components on VisA and MVTec AD Datasets. Our contributions are **bold**.

4.4. Ablation Study

Ablation Study on Different Components. Our approach comprises two key components: the decoupled alignment framework and the DAR block. To validate the efficacy and

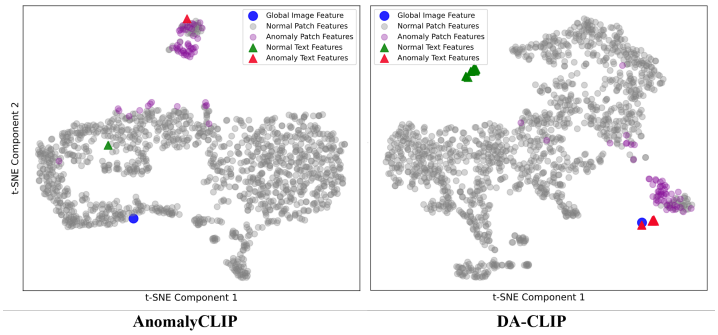


Figure 6. Visualization of the embedding spaces for DA-CLIP and AnomalyCLIP. AnomalyCLIP's text features neglect to align with global image features, while DA-CLIP achieves precise alignment.

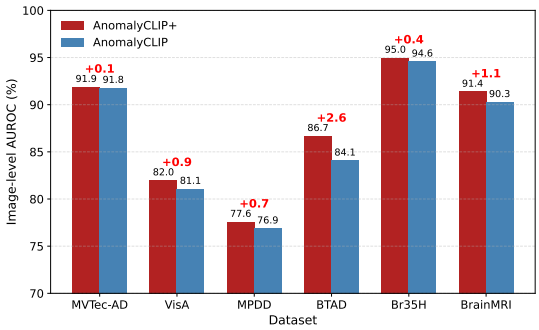


Figure 7. Absolute Image-level AUROC Improvement of using Token Adapter (AnomalyCLIP+).

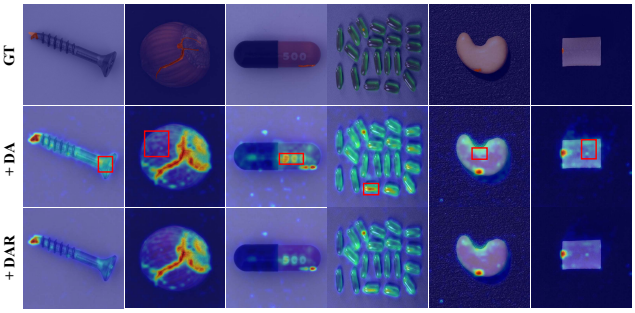


Figure 8. Visualization of ablation study on different components. (DA: decoupled alignment. DAR: decoupled asymmetric refinement block)

necessity of these modules, we conducted ablation studies on the MVTec AD and VisA datasets, as summarized in Tab. 3. While joint Alignment enhances segmentation, it degrades classification below original CLIP performance. In contrast, our decoupled alignment resolves this limitation, delivering significant gains in both classification and segmentation performance. Specifically, we observe absolute gains of 18.2% in image-level AUROC and 15.6% in AP on VisA dataset. Finally, when all components are combined, we observe the best overall performance. Fig. 8 provides an intuitive illustration of each design's contribution.

Ablation Study on Decoupled Asymmetric Refinement Block. Tab. 4 demonstrates the effectiveness of our DAR block's key design principles. Undecoupled text embeddings cause significant performance degradation, validating the necessity of decoupled alignment. Moreover, the asymmetric architecture reduces the risk of overfitting on the auxiliary dataset, and the refinement module leverages visual priors to dynamically adjust text embeddings. The results indicate that combining all three components yields the best overall performance.

Ablation Study on Different Text Refinement. Using fine-grained image features to refine text embeddings is a common practice. As demonstrated in Tab. 5, the results clearly demonstrate that the framework combining self-attention with Mamba outperforms mainstream cross-attention mechanism.

Method	Pixel-Level		Image-Level	
	AUROC	PRO	AUROC	AP
DAR	**91.2**	**87.3**	**91.9**	**96.2**
w/o Decoupled	91.1(-0.1)	87.1(-0.2)	90.1(-1.8)	95.5(-0.7)
w/o Asymmetric	90.6(-0.6)	84.8(-2.5)	91.5(-0.4)	96.2(-0.0)
w/o Refinement	91.0(-0.2)	87.0(-0.3)	91.2(-0.7)	95.8(-0.4)

Table 4. Ablation study on core designs of the DAR block.

Method	Pixel-Level		Image-Level	
	AUROC	PRO	AUROC	AP
Cross Attention	91.1(-0.1)	87.0(-0.3)	90.5(-1.4)	95.9(-0.3)
Hybrid Attention	**91.2**	**87.3**	**91.9**	**96.2**

Table 5. Ablation study on different text refinement.

5. Conclusion

In this work, we analyze a fundamental yet overlooked the granularity mismatch problem in ZSAD from the perspective of cross-modal alignment. We observe that in a unified embedding space, text features preferentially align with local visual features, which in turn widens the modality gap with global visual features. This leads to a situation where CLIP can locate but does not recognize anomalies. To address this issue, we propose DA-CLIP, which successfully mitigates the alignment bias caused by granularity differences by separating the text feature alignment process for classification and segmentation. In addition, the DAR block dynamically adjusts text features by incorporating visual priors, achieving a powerful alignment. Extensive experiments conducted on 11 industrial and medical datasets demonstrate the superior performance of DA-CLIP.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China (No.62372082 and No.62572100), the Fundamental Research Funds for the Central Universities (No.ZYGX2024Z017), Shenzhen Natural Science Foundation (No.JCYJ20240813114206010), and Graduate Teaching Research and Reform Project of the Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China (No.JYJG2026007).

References

[1] Javeria Amin, Muhammad Sharif, Mussarat Yasmin, and Steven Lawrence Fernandes. A distinctive approach in brain tumor detection and classification using mri. *Pattern Recognition Letters*, 139:118–127, 2020. 6

[2] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 6

[3] Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5564–5572, 2023. 6

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 6

[5] 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

[6] 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, 2024. 1, 3, 5

[7] Xuhai Chen, Yue Han, and Jiangning Zhang. 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*, 2(4), 2023. 1, 3, 6, 7

[8] Xuhai Chen, Jiangning Zhang, Guanzhong Tian, Haoyang He, Wuhao Zhang, Yabiao Wang, Chengjie Wang, and Yong Liu. Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection. In *International Joint Conference on Artificial Intelligence*, pages 17–33. Springer, 2024. 1, 3

[9] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018. 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. 2

[11] Bin-Bin Gao, Yue Zhou, Jiangtao Yan, Yuezhi Cai, Weixi Zhang, Meng Wang, Jun Liu, Yong Liu, Lei Wang, and Chengjie Wang. Adaptclip: Adapting clip for universal visual anomaly detection. *arXiv preprint arXiv:2505.09926*, 2025. 1, 4

[12] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Hao Li, Ming Tang, and Jinqiao Wang. Filo: Zero-shot anomaly detection by fine-grained description and high-quality localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2041–2049, 2024. 3, 5, 6

[13] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25261–25270, 2025. 6

[14] Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 37:71162–71187, 2024. 2

[15] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8472–8480, 2024.

[16] Liren He, Zhengkai Jiang, Jinlong Peng, Wenbing Zhu, Liang Liu, Qiangang Du, Xiaobin Hu, Mingmin Chi, Yabiao Wang, and Chengjie Wang. Learning unified reference representation for unsupervised multi-class anomaly detection. In *European Conference on Computer Vision*, pages 216–232. Springer, 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. 1

[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. 1, 3, 6, 7

[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] Gahyeon Kim, Sohee Kim, and Seokju Lee. Aapl: Adding attributes to prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1572–1582, 2024. 1

[21] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9664–9674, 2021. 2

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[23] Xinyue Liu, Jianyuan Wang, Biao Leng, and Shuo Zhang. Unlocking the potential of reverse distillation for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5640–5648, 2025. 2

[24] Jitao Ma, Weiying Xie, Hangyu Ye, Daixun Li, and Leyuan Fang. Aligning and prompting anything for zero-shot generalized anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5964–5972, 2025. 6, 7

[25] Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4744–4754, 2025. 1, 3, 6

[26] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 6

[27] 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

[28] Zhen Qu, Xian Tao, Mukesh Prasad, Fei Shen, Zhengtao Zhang, Xinyi Gong, and Guiguang Ding. Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. In *European Conference on Computer Vision*, pages 301–317. Springer, 2024. 1, 4, 5

[29] Zhen Qu, Xian Tao, Xinyi Gong, Shichen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30398–30408, 2025. 1, 3, 6, 7

[30] 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. 1

[31] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912, 2021. 6

[32] Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *European Conference on Computer Vision*, pages 474–489. Springer, 2022. 2

[33] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan Duong, Chanh D Tr Nguyen, and Steven QH Truong. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24511–24520, 2023. 2

[34] Hantao Yao, Rui Zhang, and Changsheng Xu. Tcp: Textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23438–23448, 2024. 1

[35] 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

[36] 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

[37] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draema discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8330–8339, 2021. 2

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

[39] Ximiao Zhang, Min Xu, Dehui Qiu, Ruixin Yan, Ning Lang, and Xiuzhuang Zhou. Mediclip: Adapting clip for few-shot medical image anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 458–468. Springer, 2024. 1

[40] Ying Zhao. Omnial: A unified cnn framework for unsupervised anomaly localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3924–3933, 2023. 2

[41] 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. 1, 3, 5, 6, 7

[42] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pretraining for anomaly detection and segmentation. In *European conference on computer vision*, pages 392–408. Springer, 2022. 6