This CVPR 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.

FB-CLIP: Fine-Grained Zero-Shot Anomaly Detection with Foreground-Background Disentanglement

Ming Hu[1,2] Yongsheng Huo[1,2] Mingyu Dou[1,2] Jianfu Yin[1,2] Peng Zhao[1,2]
Yao Wang[3] Cong Hu[4*] Bingliang Hu[1*] Quan Wang[1*]
[1] Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences [3] Xi'an Jiaotong University
[4] Zhongnan Hospital of Wuhan University

Abstract

Fine-grained anomaly detection is crucial in industrial and medical applications, but labeled anomalies are often scarce, making zero-shot detection challenging. While vision-language models like CLIP offer promising solutions, they struggle with foreground-background feature entanglement and coarse textual semantics. We propose FB-CLIP, a framework that enhances anomaly localization via multi-strategy textual representations and foreground-background separation. In the textual modality, it combines End-of-Text features, global-pooled representations, and attention-weighted token features for richer semantic cues. In the visual modality, multi-view soft separation along identity, semantic, and spatial dimensions, together with background suppression, reduces interference and improves discriminability. Semantic Consistency Regularization (SCR) aligns image features with normal and abnormal textual prototypes, suppressing uncertain matches and enlarging semantic gaps. Experiments show that FB-CLIP effectively distinguishes anomalies from complex backgrounds, achieving accurate fine-grained anomaly detection and localization under zero-shot settings. Code : https://github.com/Xi-Mu-Yu/FB-CLIP.

1. Introduction

Anomaly detection (AD) aims to identify data samples or regions that deviate significantly from normal patterns, and plays a critical role in applications ranging from industrial inspection [2, 3, 5, 9, 10, 23, 32, 35, 36, 40, 41, 50, 52, 57, 58] to medical imaging [10, 12, 31, 36, 37, 45, 46]. Despite its importance, labeled anomalous data are often scarce or unavailable [22], motivating research into unsupervised and zero-shot anomaly detection (ZSAD) [24, 58]. Recent advances in vision–language pretraining models (VLMs) [27, 39], particularly CLIP [39], have enabled semantic alignment between visual and textual modalities, opening new opportunities for ZSAD.

Although CLIP has opened a promising direction for ZSAD, directly applying it to fine-grained anomaly detection remains challenging. By examining the visualization of the original CLIP responses presented in AnomalyCLIP (see Fig. 2 (a)), we observe CLIP often produces strong responses in both foreground and background regions si-

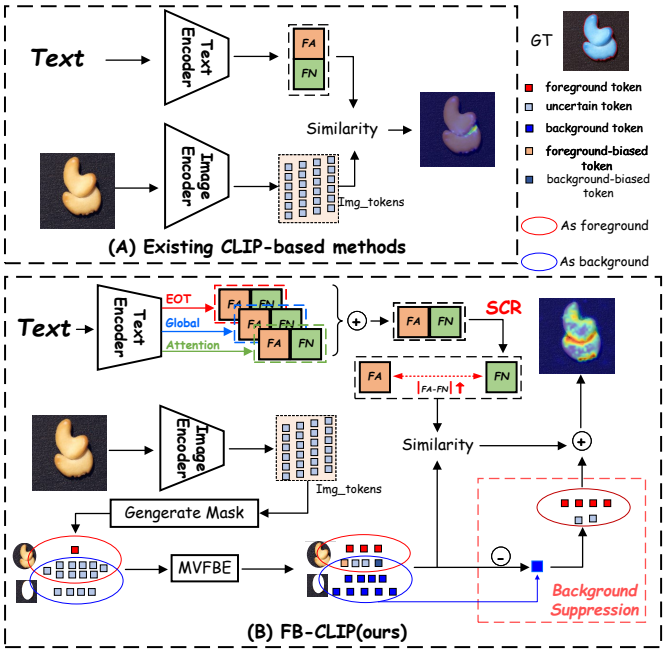


Figure 1. Framework comparison. (a) Existing methods use a single text feature for uniform alignment, leading to coarse semantics. (b) FB-CLIP employs multi-text features and Semantic Consistency Regularization (SCR) to enhance alignment, while foreground–background separation and background suppression enable fine-grained, noise-robust anomaly representation.

*Corresponding author.

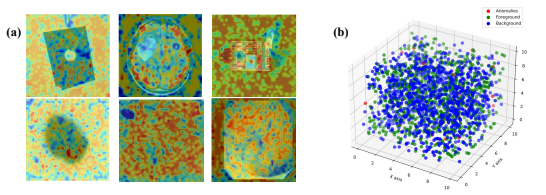


Figure 2. (a) Visualization of the responses from the original CLIP model, corresponding to Fig. 1(c) in AnomalyCLIP [58]. (b) Entanglement of foreground, background, and anomalous regions. Separating foreground and background clarifies the distinction between anomalous and normal points, improving accuracy.

multaneously. This indicates that the model struggles to distinguish anomaly-related foreground semantics from irrelevant background context, resulting in pronounced foreground–background feature entanglement in visual representations [17, 30, 51], as illustrated in Fig. 2 (b). In such cases, abundant background responses may obscure subtle anomaly signals, making it difficult for the model to accurately perceive and localize fine-grained anomalies.

As illustrated in Fig. 1(a), existing CLIP-based anomaly detection methods attempt to alleviate this issue mainly from two perspectives: text modeling and visual feature enhancement. On the textual side [6, 11, 24, 38, 58, 59], many approaches introduce learnable prompts or manually designed descriptions to project textual semantics into the shared embedding space and align them with image features. Representative works such as AnomalyCLIP attempt to learn object-agnostic prompts that encode generic normal and abnormal semantics. While these strategies improve cross-domain generalization to some extent, the textual representations themselves often remain relatively coarse and limited in semantic diversity. As a result, the textual features may provide insufficient guidance for distinguishing foreground anomaly cues from complex background contexts. In fine-grained anomaly scenarios, where abnormal patterns are often subtle and sparsely distributed, such limited semantic expressiveness can lead to ambiguous cross-modal alignment, making it difficult for the model to clearly separate anomaly-related regions from irrelevant background responses. On the visual side, existing methods typically enhance image tokens to improve patch-level feature representation and alignment [6, 11, 13, 24]. However, these approaches usually treat tokens in a homogeneous manner, implicitly assuming that all tokens contribute equally to anomaly representation. This assumption overlooks the large uncertainty and diversity of semantic information across different image regions. In fine-grained anomaly scenarios, some tokens may contain critical anomaly cues, while others mainly capture background context or noise. Without explicitly modeling such differences, the enhanced representations may still suffer from foreground–background ambiguity, which dilutes anomaly responses and reduces localization accuracy.

To address the foreground–background ambiguity in CLIP-based anomaly detection, as illustrated in Fig. 1(b), we propose FB-CLIP, a framework designed to improve anomaly localization through foreground–background separation guided by richer semantic representations. In the textual modality, FB-CLIP constructs more expressive textual features by integrating End-of-Text (EOT) features, global-pooled representations, and attention-weighted token features. Compared with conventional prompt-based text representations, this multi-strategy feature fusion provides more comprehensive semantic cues about normality and abnormality. The richer textual semantics serve as stronger guidance for cross-modal alignment, enabling the model to better differentiate anomaly-related foreground signals from background context during feature matching. To further stabilize cross-modal alignment and improve fine-grained anomaly localization under zero-shot settings, we introduce Semantic Consistency Regularization (SCR). SCR aligns image features with both normal and abnormal textual prototypes, and by minimizing prediction entropy and imposing margin constraints between semantic prototypes, it suppresses uncertain matches and enlarges semantic gaps. This enhances semantic discriminability within the single-text-space alignment, further improving anomaly localization performance. In the **visual modality**, FB-CLIP performs multi-view foreground–background soft separation along *identity*, *semantic*, and *spatial* dimensions to refine token representations. Continuous gating integrates dual-branch attention outputs, preventing information loss caused by hard filtering. Moreover, the Background Suppression (BS) module leverages multi-layer, multi-scale background prototypes to further reduce residual interference, mitigating the impact of complex scenes and cross-domain distribution shifts. These designs enable FB-CLIP to produce cleaner and more discriminative anomaly representations, facilitating accurate detection and localization. Our contributions are as follows:

Multi-strategy text representation. We construct semantically rich and task-aware textual representations by integrating EOT, global-pooled, and attention-weighted token features, enabling more robust anomaly perception.

Enhanced visual representation. We propose a multi-view foreground–background soft separation strategy along identity, semantic, and spatial dimensions, combined with background suppression, to capture fine-grained visual cues while reducing background noise.

Cross-modal discriminative alignment. We introduce SCR to align image features with normal and abnormal textual prototypes, suppress uncertain matches, and enlarge similarity gaps, achieving stable cross-modal alignment.

2. Related work

2.1. Prompt Design and Textual Adaptation

In CLIP-based adaptation, prompt engineering plays a key role. Early methods such as CoOp [56] and CoCoOp [55] learn context tokens for flexible textual adaptation, while AnomalyCLIP [58] and PromptAD [29] use learnable prompts to capture generic or task-specific anomaly semantics. However, these approaches rely solely on EOT sentence embeddings, potentially overlooking token-level context. To address this, we propose **MSTFF**, which combines EOT, global pooling, and attention-based pooling to generate task-aware text representations that enhance anomaly sensitivity. Additionally, **SCR** enforces confidence and discriminative constraints on visual-text alignment, improving both text representations and ZSAD performance.

2.2. Visual Feature Enhancement

In terms of visual features, several works have attempted to enhance CLIP's spatial representations to improve anomaly detection. For example, WinCLIP [24] achieves dense anomaly mapping via sliding-window patch evaluation, CLIP-AD [8] utilizes local patch–text similarities for anomaly localization, VisionAD [44] combines hierarchical features from multiple transformer layers, and AF-CLIP [11] employs neighborhood aggregation for feature enhancement. However, these methods primarily operate at the token level and overlook the semantic and structural information within each token. To address this, we propose the **MVFBE** and **BS** modules to disentangle and enhance semantic and spatial features, highlighting anomalous signals while suppressing background redundancy.

3. Method

3.1. Multi-Strategy Text Feature Fusion (MSTFF)

Following **AnomalyCLIP** [58], we adopt **object-agnostic learnable text prompts** for ZSAD, which discard class-specific semantics to capture general anomaly patterns. The normal and abnormal prompts are defined as:

$$g_n = [V_1][V_2]\dots[V_E][\text{object}], \tag{1}$$
$$g_a = [W_1][W_2]\dots[W_E][\text{damaged}][\text{object}], \tag{2}$$

where $[V]_i$ and $[W]_i$ denote learnable embeddings for normal and abnormal prompts, respectively. This object-agnostic formulation directs the model to focus on anomaly-related cues rather than object categories, enabling better cross-object and cross-domain generalization.

CLIP's text encoder typically relies on the End-of-Text (EOT) token for sentence-level representation, which may overlook contextual or anomaly-relevant semantics. To enhance textual representation, we propose a **Multi-Strategy Text Feature Fusion (MSTFF)** (as shown in Figure 3(a)) module that combines complementary pooling strategies for richer and more discriminative text features .

Given the transformer-encoded token sequence $\mathbf{X} \in \mathbb{R}^{B\times L\times D}$, MSTFF extracts three types of features:
EOT feature for maintaining CLIP compatibility:

$$\mathcal{F}_{\text{eot}} = \mathbf{X}[\text{arange}(B), \text{argmax}(\mathbf{t}_{\text{token}})]\mathbf{W}_{\text{proj}}. \tag{3}$$

Global feature via mean pooling to capture context:

$$\mathcal{F}_{\text{global}} = \frac{1}{L}\sum_{i=1}^{L}\mathbf{X}[:, i, :]\mathbf{W}_{\text{proj}}. \tag{4}$$

Attention feature guided by a lightweight token selector $\mathcal{S}(\cdot)$ implemented as a two-layer MLP:

$$\mathcal{F}_{\text{attn}} = \sum_{i=1}^{L}\text{softmax}\big(\mathcal{S}(\mathbf{X}[:, i, :])\big)\mathbf{X}[:, i, :]. \tag{5}$$

Final text representation by weighted fusion:

$$\mathcal{F}_{\text{text}} = \lambda_1\mathcal{F}_{\text{global}} + \lambda_2\mathcal{F}^{\text{proj}}_{\text{attn}} + \lambda_3\mathcal{F}_{\text{eot}}, \tag{6}$$

where $\lambda_1 = 1.0$, $\lambda_2 = \lambda_3 = 0.5$.

Through this fusion, MSTFF captures complementary semantics: the EOT feature preserves CLIP alignment, the global feature provides contextual stability, and the attention feature highlights task-relevant tokens. This design yields expressive, adaptive, and generalizable text representations for zero-shot anomaly detection.

3.2. Multi-View Foreground-Background Enhancement (MVFBE)

While MSTFF focuses on textual representations, visual features still suffer from foreground-background ambiguity. To address this challenge, we propose a novel **Multi-View Foreground-Background Enhancement (MVFBE)** (as shown in Figure 3(b)) module that explicitly separates and enhances foreground and background features from complementary perspectives. Our key insight is that foreground tokens, which typically contain anomalies, exhibit higher *information richness* (diversity from the global context), while background tokens demonstrate greater *stability* (consistency with normal patterns).

Given input visual tokens $\mathbf{X} \in \mathbb{R}^{B\times(L+1)\times C}$ extracted from a pre-trained Vision Transformer (ViT), including one class token $\mathbf{x}_{\text{cls}} \in \mathbb{R}^{B\times 1\times C}$ and L patch tokens $\mathbf{x}_{\text{img}} \in \mathbb{R}^{B\times L\times C}$, the MVFBE module first generates a soft foreground probability mask $\mathbf{P}_{\text{fg}} \in \{0.5, 1.0\}^{B\times L}$ using multiple complementary anomaly indicators. Specifically, we compute four complementary cues: (1) *local saliency* $\mathcal{S}_{\text{local}}$ (2) *center distance* $\mathcal{D}_{\text{center}}$ (3) *CLS inconsistency* $\mathcal{I}_{\text{cls}}$ and (4) *temporal variation* $\mathcal{T}_{\text{temp}}$. Details in Appendix A.

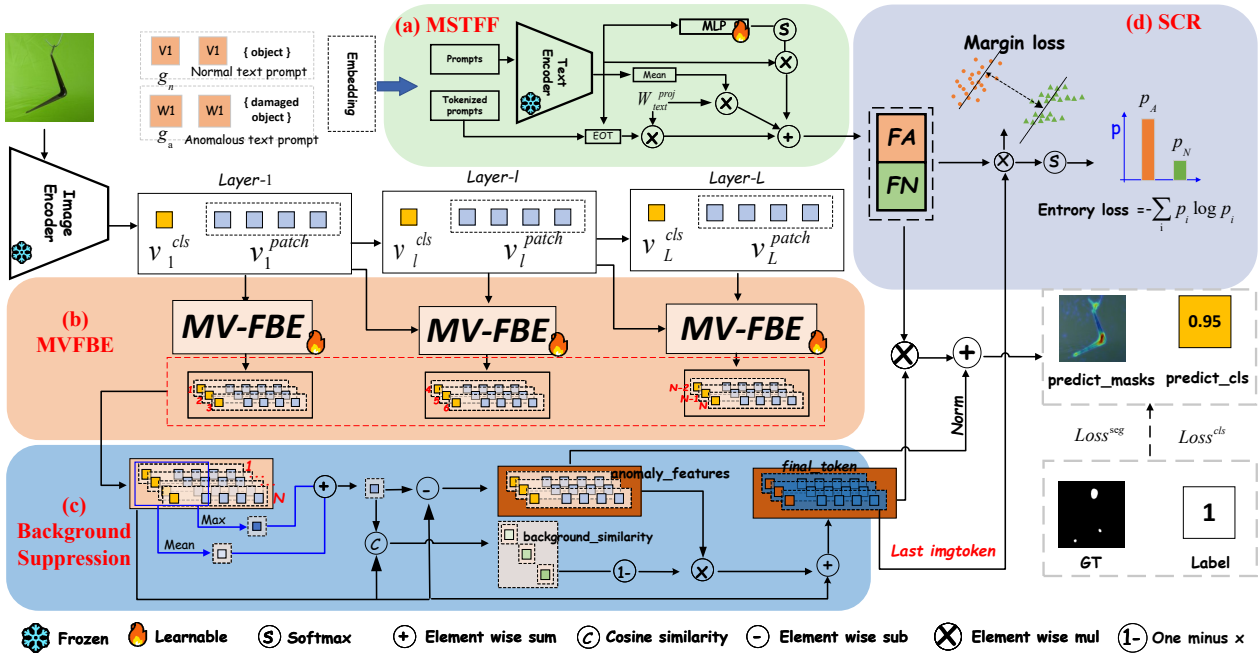


Figure 3. Overview of **FB-CLIP** for zero-shot anomaly detection. (a) Multi-Strategy Text Feature Fusion (MSTFF) generates task-aware text embeddings (see Section3.1); (b) Multi-View Foreground-Background Enhancement (MVFBE) separates and enhances foreground and background features (see Section3.2); (c) Background Suppression reduces residual background interference (see Section3.3); (d) Semantic Consistency Regularization (SCR) enforces confident and discriminative visual-text alignment. These components jointly enable fine-grained anomaly detection without labeled abnormal samples (see Section3.4).

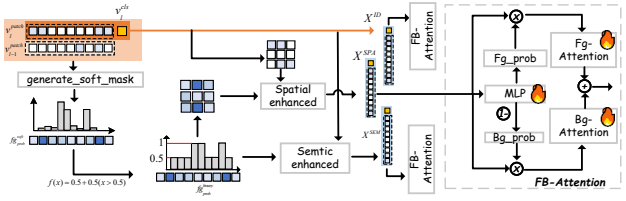


Figure 4. Structure of the proposed **MVFBE** module. The module first generates a soft foreground mask from multiple anomaly indicators. Three complementary views are applied: Identity (ID) preserves original features, Semantic (SEM) models foreground richness and background stability, and Spatial (SPA) captures fine-grained local structures. Features are then integrated across layers and refined with foreground-background attention to produce discriminative anomaly representations. Details in Appendix A, **??**

.

The normalized indicators are then linearly combined to form an anomaly score:

$$\mathcal{A} = \alpha_1 \tilde{\mathcal{S}}_{\text{local}} + \alpha_2 \tilde{\mathcal{D}}_{\text{center}} + \alpha_3 \tilde{\mathcal{I}}_{\text{cls}} + \alpha_4 \tilde{\mathcal{T}}_{\text{temp}}, \quad (7)$$

where $\alpha_1 = \alpha_2 = \alpha_3 = 0.3$ and $\alpha_4 = 0.1$ to emphasize spatial and semantic cues while incorporating temporal information when available. The anomaly score is then binarized to generate the soft foreground mask:

$$\mathbf{P}_{\text{fg}}[i] = \begin{cases} 1.0 & \text{if } \mathcal{A}[i] > 0.5 \\ 0.5 & \text{otherwise} \end{cases}, \quad (8)$$

where the threshold of 0.5 is chosen based on the per-sample normalized score distribution. The soft mask design with values $\{0.5, 1.0\}$ maintains gradient stability during backpropagation while ensuring clear separation: tokens with $\mathbf{P}_{\text{fg}}[i] = 1.0$ represent high-confidence foreground regions, while those with $\mathbf{P}_{\text{fg}}[i] = 0.5$ are treated as uncertain/background tokens that require further refinement in subsequent enhancement stages.

3.2.1. Three Complementary Enhancement Views

The MVFBE module applies three complementary enhancement views to process the visual tokens.

Identity View (ID). The **Identity (ID)** view preserves original features:

$$\mathbf{X}_{\text{ID}} = \mathbf{X}. \quad (9)$$

This serves as a baseline to retain raw feature information and prevent degradation from over-transformation.

Semantic View (SEM). The **SEM** view performs global self-attention-based enhancement by explicitly modeling *information richness* for foreground tokens and *stability* for background tokens. (The pseudocode in Appendix **??**)

We first construct interaction weight matrices:

$$\mathbf{W}_{\text{fg}} = \mathbf{p}_{\text{fg}} \otimes \mathbf{p}_{\text{fg}}^T, \quad \mathbf{W}_{\text{bg}} = (1 - \mathbf{p}_{\text{fg}}) \otimes (1 - \mathbf{p}_{\text{fg}})^T, \tag{10}$$

where $\otimes$ denotes outer product. Foreground information richness is computed by deviation from the class token:

$$\mathbf{r}_{\text{info}} = 1 - \text{cosine}(\mathbf{X}_{\text{tokens}}, \mathbf{X}_{\text{cls}}), \tag{11}$$

and attention weights are:

$$\mathbf{A}_{\text{fg}} = \text{Softmax}(\mathbf{r}_{\text{info}} \odot \mathbf{W}_{\text{fg}}). \tag{12}$$

For background tokens, stability is measured by similarity:

$$\begin{aligned} \mathbf{s}_{\text{stable}} &= \text{cosine}(\mathbf{X}_{\text{tokens}}, \mathbf{X}_{\text{cls}}), \\ \mathbf{A}_{\text{bg}} &= \text{Softmax}(\mathbf{s}_{\text{stable}} \odot \mathbf{W}_{\text{bg}}). \end{aligned} \tag{13}$$

Foreground and background features are aggregated:

$$\mathbf{X}_{\text{fg,agg}} = \mathbf{A}_{\text{fg}}\mathbf{X}_{\text{tokens}}, \quad \mathbf{X}_{\text{bg,agg}} = \mathbf{A}_{\text{bg}}\mathbf{X}_{\text{tokens}}, \tag{14}$$

and fused with a residual connection:

$$\begin{aligned} \mathbf{X}_{\text{SEM}} = \alpha[\mathbf{p}_{\text{fg}} \odot \mathbf{X}_{\text{fg,agg}} + (1 - \mathbf{p}_{\text{fg}}) \odot \mathbf{X}_{\text{bg,agg}}] \\ + (1 - \alpha)\mathbf{X}_{\text{tokens}}, \alpha = 0.6. \end{aligned} \tag{15}$$

Spatial View (SPA). Inspired by PatchCore [41], MUSC [28] , AF-CLIP [11], CLIP-Haze [54], and MFP-CLIP [53], we design a spatial enhancement module from the perspective of neighborhood aggregation to more effectively capture local contextual information. Unlike existing methods, we explicitly utilize the unfolded mask to distinguish foreground and background patches, and further process them based on foreground information richness, introducing a fundamentally different spatial modeling paradigm.

The **Spatial (SPA)** view performs local patch-based enhancement to capture fine-grained spatial structures. Tokens are reshaped into 2D grids and unfolded into 5×5 overlapping patches (The pseudocode in Appendix **??**):

$$\mathbf{X}_{\text{unfold}} \in \mathbb{R}^{(BHW)\times 25\times C}. \tag{16}$$

Foreground and background patches are separated using the unfolded mask. Background stability combines cosine similarity with feature magnitude:

$$\mathbf{s}_{\text{bg}} = \text{cosine}(\mathbf{X}_{\text{bg,patches}}, \mathbf{X}_{\text{cls}}) \odot \|\mathbf{X}_{\text{bg,patches}}\|, \tag{17}$$

while foreground information richness considers deviation and local diversity:

$$\begin{aligned} \mathbf{r}_{\text{fg}} =& (1 - \text{cosine}(\mathbf{X}_{\text{fg,patches}}, \mathbf{X}_{\text{cls}})) \\ & \odot \|\mathbf{X}_{\text{fg,patches}} - \mu(\mathbf{X}_{\text{fg,patches}})\|. \end{aligned} \tag{18}$$

Weighted aggregation produces the enhanced features:

$$\mathbf{X}_{\text{fg/bg,agg}} = \frac{\sum_{i=1}^{25} \text{Softmax}(\mathbf{w}_{\text{fg/bg}})[i] \odot \mathbf{X}_{\text{fg/bg,patches}}[i]}{\sum_{i=1}^{25} \mathbf{P}_{\text{fg/bg,unfold}}[i]}, \tag{19}$$

$$\mathbf{X}_{\text{SPA}} = \text{Reshape}(\mathbf{X}_{\text{fg,agg}} + \mathbf{X}_{\text{bg,agg}}). \tag{20}$$

3.2.2. Multi-Layer Integration

The three enhanced views form:

$$\mathbf{F}_{\text{enhanced}} = [\mathbf{X}_{\text{ID}}, \mathbf{X}_{\text{SEM}}, \mathbf{X}_{\text{SPA}}], \tag{21}$$

and are applied across N Transformer layers, producing $3N$ enhanced feature groups. Inspired by AF-CLIP [11], each feature group is further processed by the **Foreground-Background Attention (FB-Attention)** (shown in Fig 4) module with adaptive gating (inspired by LSTM [20, 21]):

$$\mathbf{x}_{\text{fg},i} = \mathcal{G}(\mathbf{x}_i)\mathbf{x}_i, \quad \mathbf{x}_{\text{bg},i} = (1 - \mathcal{G}(\mathbf{x}_i))\mathbf{x}_i, \tag{22}$$

where $\mathcal{G} : \mathbb{R}^C \to [0, 1]$ is a learnable gate function. Apply multi-head self-attention (MHSA) [47] to each stream separately, then sum the outputs and apply LayerNorm (LN).

$$\mathbf{H}_{\text{FB}} = \text{LN}(\text{MHSA}_{\text{fg}}(\mathbf{X}_{\text{fg}}) + \text{MHSA}_{\text{bg}}(\mathbf{X}_{\text{bg}})). \tag{23}$$

3.3. Background Suppression (BS)

After MVFBE, we address the critical challenge of **background interference** in ZSAD. Although the enhancement module separates semantic regions, background tokens may still retain residual noise that degrades detection accuracy. To alleviate this, we introduce a **background subtraction mechanism** (as shown in Figure 3(c)) that explicitly models and suppresses background patterns while preserving anomaly-specific signals.

Given the $3N$ enhanced features $\mathbf{F}_{\text{enhanced}}$, we extract a prototypical background representation by aggregating the first half of candidate background tokens from each feature.

$$\mathbf{X}_{\text{bg,cand}}^{(i)} = \mathbf{F}_{\text{enhanced}}^{(i)}[:, : L/2, :]. \tag{24}$$

These candidates are concatenated to form a unified background feature bank:

$$\mathbf{X}_{\text{bg,bank}} = \text{Concat}\big[\mathbf{X}_{\text{bg,cand}}^{(1)}, \ldots, \mathbf{X}_{\text{bg,cand}}^{(3N)}\big], \tag{25}$$

where $\mathbf{X}_{\text{bg,bank}} \in \mathbb{R}^{B\times M\times C}$ and M is the total number of background tokens. The prototypical background vector is computed as:

$$\mathbf{b}_{\text{proto}} = \tfrac{1}{2}\,\text{Mean}(\mathbf{X}_{\text{bg,bank}}) + \tfrac{1}{2}\,\text{Max}(\mathbf{X}_{\text{bg,bank}}), \tag{26}$$

which combines global averaging and max-pooling to preserve salient background characteristics.

For each feature $\mathbf{F}_{\text{enhanced}}^{(i)} \in \mathbb{R}^{B\times L\times C}$, we compute cosine similarity to the background prototype:

$$\mathbf{s}_{\text{bg}}^{(i)} = \text{cosine}\Big(\mathbf{F}_{\text{enhanced}}^{(i)}, \mathbf{b}_{\text{proto}}\Big) \in \mathbb{R}^{B\times L\times 1}. \tag{27}$$

The anomaly signal is then obtained by background subtraction:

$$\mathbf{a}^{(i)} = \mathbf{F}_{\text{enhanced}}^{(i)} - \mathbf{b}_{\text{proto}}, \tag{28}$$

and its reconstruction error is defined as:

$$\mathbf{e}^{(i)} = \|\mathbf{a}^{(i)}\|_2. \tag{29}$$

To suppress residual background while emphasizing anomaly cues, we apply similarity-weighted enhancement:

$$\mathbf{a}^{(i)}_{\text{enh}} = \mathbf{a}^{(i)} \odot (1 - \mathbf{s}^{(i)}_{\text{bg}}), \tag{30}$$

where tokens highly similar to the background receive smaller weights. The final background-suppressed features:

$$\mathbf{F}^{(i)}_{\text{final}} = \alpha\, \mathbf{F}^{(i)}_{\text{enhanced}} + (1 - \alpha)\, \mathbf{a}^{(i)}_{\text{enh}}, \quad \alpha = 0.5. \tag{31}$$

3.4. Semantic Consistency Regularization (SCR)

Even after visual disentanglement, visual-text alignment may remain unstable in the zero-shot setting. Therefore, inspired by Support Vector Machines [18], we introduce a **semantic consistency regularization** (as shown in Figure 3(d)) loss that jointly enforces *confidence* and *discriminative* constraints on visual–text alignment without additional supervision. Given image token $\mathbf{V}_{\text{tokens}} \in \mathbb{R}^{B\times N\times D}$ (from the last layer after background suppression, representing the most semantically rich anomaly-focused representations) and text features $\mathbf{T} \in \mathbb{R}^{2\times D}$ (for "normal" and "abnormal" prompts), we aggregate token-level features via global average pooling and apply L2 normalization:

$$\mathbf{V} \leftarrow \text{L2Norm}(\mathbf{V}), \quad \mathbf{T} \leftarrow \text{L2Norm}(\mathbf{T}). \tag{32}$$

Semantic similarity is computed as:

$$\mathbf{s} = \tfrac{1}{\tau}\mathbf{V}\mathbf{T}^\top, \quad \mathbf{s} \in \mathbb{R}^{B\times 2},\ \tau = 0.07, \tag{33}$$

where $\mathbf{s}[b, 0]$ and $\mathbf{s}[b, 1]$ denote similarities to normal and abnormal prompts, respectively. Class probabilities are:

$$\mathbf{p}_b = \text{Softmax}(\mathbf{s}_b), \quad \mathbf{p}_b \in \mathbb{R}^2, \tag{34}$$

Entropy Regularization. We encourage confident predictions by minimizing entropy:

$$\mathcal{L}_{\text{entropy}} = \mathbb{E}^B_{b=1}\left[-\sum_{c=0}^{1} p_b(c)\log\left(p_b(c)\right)\right] \tag{35}$$

Margin Regularization. We enforce discriminative separation via margin constraint:

$$\mathcal{L}_{\text{margin}} = \mathbb{E}^B_{b=1}[\max(0,\, \gamma - |s_b[1] - s_b[0]|)]\,, \tag{36}$$

where $\gamma = 1$ is the margin parameter. This prevents mode collapse and maintains clear separation between normal and abnormal semantics. The final loss combines both terms:

$$\mathcal{L}_{\text{consistency}} = \lambda\big(w_e\, \mathcal{L}_{\text{entropy}} + w_m\, \mathcal{L}_{\text{margin}}\big), \tag{37}$$

where $\lambda = 0.15$ controls the overall weight, and $w_e = 1.0$, $w_m = 0.5$ are relative coefficients for the entropy and margin terms. This dual regularization strategy addresses two essential aspects : (1) encouraging confident and semantically consistent predictions, (2) maintaining discriminative separation between normal and abnormal semantics.

4. Experiment

4.1. Experiment setup

Datasets and Evaluation Metrics. We evaluate FB-CLIP on 16 publicly available datasets spanning industrial inspection and medical imaging. Industrial benchmarks include MVTec AD [2], VisA [60], MPDD [25], BTAD [34], DAGM [49], DTD-Synthetic [1] and a large-scale dataset Real-IAD [48], while medical datasets cover colon polyp (CVC-ClinicDB [4], CVC-ColonDB [43]), skin cancer detection dataset ISIC [15] and brain tumor (BrainMRI [42], Br35H [16]) detection, Kvasir [26], and Endo [19], thyroid nodule detection dataset TN3k [14], brain tumor detection datasets HeadCT [42]. Competing SOTA methods include WinCLIP [24], VAND [7], AnomalyCLIP [58], AdaCLIP [6], AA-CLIP [33], AF-CLIP [11], and FAPrompt [59]. Performance is measured using AUROC for detection, AP for anomaly classification, and AUPRO [3] for localization.

Implementation Details. FB-CLIP is built on the public CLIP backbone (`ViT-L/14@336px`) with frozen parameters. Following prior works [11, 24, 58], we fine-tune AnomalyCLIP using the test data on MVTec AD and evaluate the ZSAD performance on other datasets. As for MVTec AD, we fine-tune AomalyCLIP on the test data of VisA. Results are averaged over all sub-datasets. All experiments are implemented in PyTorch 2.0.0 and conducted on a single NVIDIA RTX 3090 GPU (24GB) using the Adam optimizer with a learning rate of 5e-5 and a batch size of 4.

4.2. Main results

As shown in Table 1, FB-CLIP demonstrates competitive performance in zero-shot anomaly detection across both industrial and medical domains. Compared with recent strong CLIP-based models such as AF-CLIP and FAPrompt, our method consistently achieves superior results on both image-level and pixel-level metrics in the industrial domain. FB-CLIP achieves notable improvements on the VisA dataset and attains over a 6% increase in pixel-level AUPRO on the large-scale Real-IAD dataset. While FAPrompt remains competitive on certain medical datasets, FB-CLIP ranks among the top two in 14 out of 18 metrics, surpassing FAPrompt, which ranks in the top two on 10 metrics.

Table 1. Zero-shot anomaly detection performance on industrial and medical domains. The image-level performance AUROC(%), AP(%) and pixel-level performance AUROC(%), PRO(%) are reported. **Red: best**, **blue: second best.**

Domain	Metric	Dataset	WinCLIP [24] CVPR'2023	VAND [7] CVPR'2023	AnomalyCLIP [58] ICLR'2024	AdaCLIP [6] ECCV'2024	AA-CLIP [33] CVPR'2025	AF-CLIP [11] MM'2025	FAPrompt [59] ICCV'2025	FB-CLIP Ours
Industrial	image-level (AUROC, AP)	MVTec	(91.8, 96.5)	(86.1, 93.5)	(91.5, 96.2)	(92.1, **96.7**)	(90.5, 94.9)	(**92.9**, **96.8**)	(91.9, 95.7)	(**92.4**,96.6)
		VisA	(78.1, 81.2)	(78.0, 81.4)	(82.1, 85.4)	(86.3, 89.2)	(84.6, 82.2)	(**88.5**, **90.0**)	(84.6, 86.8)	(**89.5**, **90.7**)
		MPDD	(63.6, 69.9)	(78.0, 81.4)	(77.0, 82.0)	(76.8, -)	(-, -)	(-, -)	(**80.1**, **83.9**)	(**79.1**, **86.2**)
		BTAD	(68.2, 70.9)	(73.5, 68.6)	(88.3, 87.3)	(92.4, 96.2)	(**93.8**, **97.9**)	(**94.3**, 95.2)	(92.2, 92.5)	(93.2, **96.4**)
		DAGM	(91.8, 79.5)	(94.4, 83.8)	(97.5, 92.3)	(98.3, 93.7)	(93.9, 84.5)	(98.7, 94.6)	(**98.8**, **95.3**)	(**99.0**, **96.5**)
		DTD	(93.2, 92.6)	(86.4, 95.0)	(93.5, 97.0)	(97.0, 99.0)	(93.3, 97.8)	(**97.9**, **99.1**)	(96.2, 98.1)	(**97.9**, **99.1**)
		Real-IAD	(-, -)	(-, -)	(-, -)	(-, -)	(-, -)	(**79.2**, **77.0**)	(77.3, 74.8)	(**80.6**, **78.4**)
	pixel-level (AUROC, PRO)	MVTec	(85.1, 64.6)	(87.6, 44.0)	(91.1, 81.4)	(86.3, 20.0)	(**91.9**, 84.6)	(**92.3**, **85.7**)	(90.6, 83.3)	(**91.9**, **85.7**)
		VisA	(79.6, 56.8)	(94.2, 86.8)	(95.5, 87.0)	(95.8, 60.0)	(95.5, 83.0)	(**96.2**, **88.7**)	(95.9, 87.7)	(**96.3**, **91.4**)
		MPDD	(76.4, 48.9)	(94.1, 83.2)	(**96.5**, 87.0)	(96.1, -)	(-, -)	(-, -)	(**96.5**, **87.9**)	(**96.9**, **91.1**)
		BTAD	(72.7, 27.3)	(89.3, 68.8)	(94.2, 74.8)	(94.0, 33.8)	(94.0, 69.0)	(94.4, **78.3**)	(**95.6**, 75.1)	(**95.8**, **80.0**)
		DAGM	(87.6, 65.7)	(82.4, 66.2)	(95.6, 91.0)	(94.5, 53.8)	(91.6, 76.5)	(97.1, 93.1)	(**98.2**, **95.0**)	(**97.3**, **94.1**)
		DTD	(83.9, 57.8)	(95.3, 86.9)	(97.9, 92.3)	(98.5, 72.9)	(96.4, 85.9)	(**98.6**, **93.8**)	(**98.3**, **93.3**)	(**98.3**, 92.6)
		Real-IAD	(-, -)	(-, -)	(-, -)	(-, -)	(-, -)	(**95.5**, 81.6)	(95.0, **82.1**)	(**95.9**, **88.2**)
Medical	image-level (AUROC, AP)	HeadCT	(81.8, 80.2)	(89.1, 89.4)	(**93.4**, 91.6)	(91.4, -)	(-, -)	(91.2, 91.1)	(**94.0**, **92.4**)	(**93.4**, **93.9**)
		BrainMRI	(86.6, 91.5)	(89.3, 90.9)	(90.3, 92.3)	(80.6, 84.4)	(91.8, 94.1)	(**95.2**, **96.3**)	(**95.8**, 96.2)	(95.1, **96.6**)
		Br35H	(80.5, 82.2)	(93.1, 92.9)	(94.6, 94.7)	(86.5, 86.3)	(89.4, 91.0)	(96.7, 96.4)	(**97.6**, **97.1**)	(**97.1**, **97.4**)
	pixel-level (AUROC, PRO)	ISIC	(83.3, 55.1)	(89.4, 77.2)	(89.7, 78.4)	(89.3, -)	(**93.9**, 87.0)	(**94.8**,**89.6**)	(91.1, 81.6)	(**93.9**, **87.7**)
		ClinicDB	(51.2, 13.8)	(80.5, 60.7)	(82.9, 67.8)	(81.6, 54.0)	(86.3, 67.1)	(**87.1**, 70.0)	(84.7, **70.1**)	(**87.2**, **73.5**)
		ColonDB	(70.3, 32.5)	(78.4, 64.6)	(81.9, 71.3)	(79.3, 49.1)	(81.6, 65.4)	(83.2, 67.9)	(**85.0**, **73.3**)	(**84.2**, **74.4**)
		Kvasir	(69.7, 24.5)	(75.0, 36.2)	(78.9, 45.6)	(77.5, 41.3)	(84.0, **50.3**)	(**84.5**, **57.5**)	(82.1, 49.9)	(**85.0**, 47.8)
		Endo	(68.2, 28.3)	(81.9, 54.9)	(84.1, 63.6)	(-, -)	(-, -)	(**88.3**, **70.3**)	(86.8, 67.6)	(**87.6**, **70.4**)
		TN3K	(70.7, 39.8)	(73.6, 37.8)	(**81.5**, **50.4**)	(77.2, -)	(-, -)	(80.5, 45.8)	(**84.7**, **54.6**)	(79.0, 46.8)

Table 2. Ablation study of different modules for zero-shot anomaly detection. Deeper colors correspond to higher average performance. The columns **MSTFF**, **MVFBE**, **BS**, and **SCR** correspond to the four improvement strategies proposed in Figure 3.

Method					MVTec		VisA	
Base	MSTFF	MVFBE	BS	SCR	Image (AUROC, AP)	Pixel (AUROC, PRO)	Image (AUROC, AP)	Pixel (AUROC, PRO)
✓					(67.1, 84.0)	(65.2, 32.9)	(45.7, 56.2)	(88.4, 71.8)
✓	✓				(82.0, 91.4)	(84.0, 68.0)	(78.3, 80.5)	(93.2, 81.4)
✓		✓			(92.2, 96.5)	(91.7, 84.5)	(88.4, 89.7)	(96.0, 90.8)
✓			✓		(44.8, 72.7)	(70.7, 38.3)	(48.3, 56.6)	(84.1, 58.3)
✓				✓	(65.4, 81.3)	(75.7, 45.5)	(63.8, 69.0)	(88.2, 65.5)
✓	✓	✓			(92.3, 96.5)	(91.2, 84.7)	(88.2, 89.7)	(95.9, 90.6)
✓	✓		✓		(73.0, 86.8)	(71.2, 28.7)	(68.3, 72.2)	(86.5, 60.2)
✓	✓			✓	(78.2, 88.8)	(83.2, 63.6)	(74.0, 74.6)	(92.4, 78.8)
✓	✓	✓	✓		**(92.8, 96.8)**	(91.6, **85.7**)	(87.6, 89.3)	(96.1, 90.6)
✓	✓	✓		✓	(92.1, 96.2)	(89.4, 78.2)	(87.9, 89.8)	(96.2, 90.7)
✓		✓	✓		(92.4, 96.7)	(91.4, 85.6)	(87.8, 89.4)	(96.1, 90.7)
✓		✓		✓	(92.1, 96.3)	(90.9, 74.6)	(88.1, 89.7)	(**96.4**, 89.3)
✓	✓	✓	✓	✓	(92.4, 96.6)	**(91.9, 85.7)**	**(89.5, 90.7)**	(96.3, **91.4**)

4.3. Ablation Study

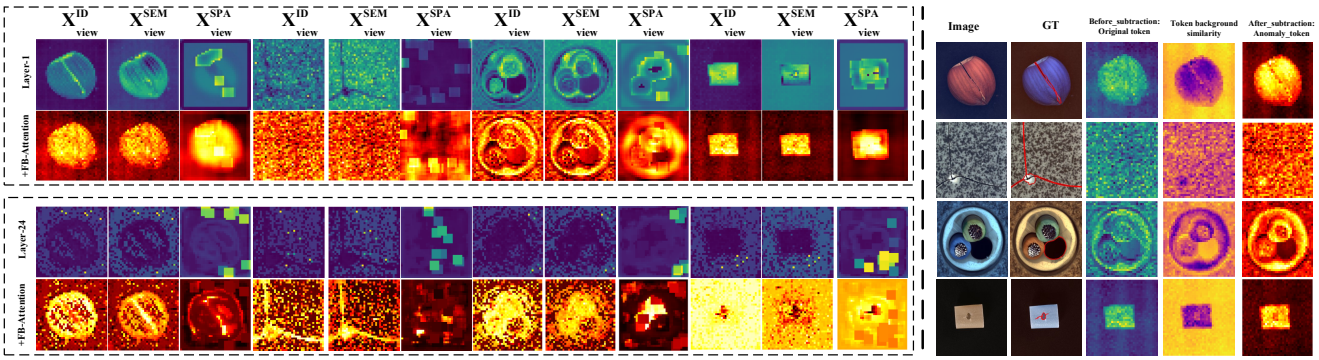
4.3.1. Effectiveness of MSTFF

The inclusion of our MSTFF module (column **MSTFF** in Table 2) leads to substantial improvements over the **Base** model. On MVTec, image-level AUROC increases from 67.1% to 82.0%, and pixel-level PRO rises from 32.9% to 68.0%, demonstrating that MSTFF effectively captures richer semantic cues and task-relevant information for more accurate anomaly recognition and localization. Additional modules—**MVFBE**, **BS**, and **SCR**—provide further gains, and the full configuration combining all four strategies achieves the best overall performance on both MVTec and VisA. These results confirm that MSTFF plays a **critical role** while synergizing with the other components. Additional ablation experiments can be found in Appendix C.

4.3.2. Effectiveness of MVFBE

The addition of the MVFBE module (column **MVFBE** in Table 2) further improves performance over the **Base** and **MSTFF** configurations. On MVTec, it increases image-level AUROC to 92.2% and pixel-level PRO to 84.5%, indicating that MVFBE effectively enhances the distinction between foreground anomalies and background context. Similar gains are observed on VisA, showing that MVFBE consistently boosts anomaly recognition and localization by emphasizing relevant object regions across multiple views.

In Figure 5(a), we visualize the effect of our multi-view imgtoken enhancement and the subsequent FB-Attention. The top row shows the first layer imgtokens obtained from the CLIP pre-trained ViT model, while the middle row displays the features from the 24th layer: $\mathbf{X}^{\text{ID}}$ along with the enhanced $\mathbf{X}^{\text{SEM}}$ and $\mathbf{X}^{\text{SPA}}$ views. Comparing the results before and after applying FB-Attention, it is evident that FB-Attention effectively suppresses background noise and enhances the foreground anomaly regions. This improvement is especially pronounced for complex textures and small defects, where the enhanced attention maps provide clearer localization and stronger contrast between anomalous objects and the background, demonstrating the benefit of integrating multi-view enhancement with FB-Attention. The block-like artifacts in the SPA view arise from noise smoothing and signal enhancement, forming connected regions, which aligns with the assumption that anomalies are connected and does not affect the capture of fine-grained local structures. Additional ablation experiments in Appendix D.



(a) Visualize the multi-view imgtoken enhancement effect and the effect after applying FB-Attention.

(b) Visualize the process of background subtraction.

Figure 5. Overall visualization of four image sets demonstrating the effects of the proposed modules. (a) Multi-View Foreground-Background Enhancement (MVFBE) results, highlighting the enhancement of tokens across multiple views. (b) Background Subtraction (BS) process, showing how irrelevant background tokens are removed to emphasize anomalies. Together, these visualizations illustrate the progressive improvement in anomaly representation and localization.

4.3.3. Effectiveness of Background Suppression

To assess the impact of token-level background subtraction, we visualize the intermediate token representations during the anomaly localization process in Figure 5(b). The first column shows the original image, the second column presents the ground truth, and the subsequent columns illustrate the background subtraction pipeline. Specifically, the "Before Subtraction" column displays the original token representations, the "Token Background Similarity" maps indicate the similarity between each token and the background tokens, and the "After Subtraction" column highlights the anomaly tokens after background removal. Darker or brighter regions in the similarity maps correspond to higher background resemblance, providing an interpretable cue for the subtraction process.

We observe that the original tokens often mix target and background information, which can obscure subtle anomalies. By leveraging background similarity, irrelevant regions are effectively suppressed, and the resulting anomaly tokens clearly emphasize defects or critical structures. Small scratches become distinctly visible, anomalies embedded in complex patterns are accurately separated, and minute anomaly points are highlighted, while noise is substantially reduced. These results demonstrate that token-level background subtraction enables the model to focus on salient regions without discarding essential features, thereby improving the clarity and reliability of anomaly localization.

4.3.4. Effectiveness of SCR

To evaluate the effectiveness of our SCR module (column **SCR** in Table 2), we conduct a detailed ablation study on both MVTec and VisA datasets. When applied individually, SCR shows limited improvement on image-level anomaly detection for MVTec (AUROC/AP: 65.4/81.3 vs. 67.1/84.0 for the Base), but substantially enhances pixel-level anomaly localization, with PRO increasing from 32.9 to 45.5. On VisA, SCR improves both image-level (AUROC/AP: 63.8/69.0 vs. 45.7/56.2) and pixel-level metrics, demonstrating its effectiveness in refining fine-grained anomaly regions. In the full configuration integrating all four strategies, it contributes to the highest image-level AUROC/AP (89.5/90.7) on VisA, outperforming all other combinations. These results indicate that SCR synergizes with the other modules to improve both global anomaly recognition and local anomaly localization, validating its role in enforcing spatial consistency and boosting fine-grained anomaly detection.

5. Conclusion

In this work, we propose **FB-CLIP**, a unified framework for zero-shot anomaly detection that achieves fine-grained and robust anomaly representations through collaborative improvements in both textual and visual modalities. In the text modality, FB-CLIP fuses EOT, global-pooled, and attention-weighted token features to construct rich, task-aware text representations. It further introduces *Semantic Consistency Regularization*, which minimizes entropy and enforces similarity margin constraints to suppress uncertain matches and enhance cross-modal alignment stability. In the visual modality, FB-CLIP leverages multi-view foreground-background soft separation and continuous gating to integrate multi-branch attention outputs, emphasizing critical anomalous cues, while *Background Suppression* reduces complex scene and cross-domain interference to produce cleaner and more robust visual representations. By integrating these strategies, FB-CLIP effectively captures fine-grained local anomaly patterns while maintaining robust ZSAD across diverse datasets and domains.

Acknowledgment

This research was supported by: The Shaanxi Province Technological Innovation Guidance Special Project: Regional Science and Technology Innovation Center, Strategic Scientific and Technological Strength Category [No.2024QY-SZX-26]. The Hubei Provincial Natural Science Foundation (Grant No. JCZRLH202600590). The research was supported by the Key Laboratory of Spectral Imaging Technology, Xi'an Institute of Optics and Precision Mechanics of the Chinese Academy of Sciences; the Key Laboratory of Biomedical Spectroscopy of Xi'an.

References

[1] 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

[2] 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. 1, 6

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020. 1, 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] Tri Cao, Jiawen Zhu, and Guansong Pang. Anomaly detection under distribution shift. *arXiv preprint arXiv:2303.13845*, 2023. 1

[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. 2, 6, 7

[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*, 2023. 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. 3

[9] Yuanhong Chen, Yu Tian, Guansong Pang, and Gustavo Carneiro. Deep one-class classification via interpolated gaussian descriptor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 383–392, 2022. 1

[10] Choubo Ding, Guansong Pang, and Chunhua Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7388–7398, 2022. 1

[11] Qingqing Fang, Wenxi Lv, and Qinliang Su. Af-clip: Zero-shot anomaly detection via anomaly-focused clip adaptation. *arXiv preprint arXiv:2507.19949*, 2025. 2, 3, 5, 6, 7

[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] 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. 2

[14] Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 257–261. IEEE, 2021. 6

[15] David Gutman, Noel C. F. Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic), 2016. 6

[16] A. Hamada. Br35h: Brain tumor detection 2020. *Online. Available: https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection*, 2020. 6

[17] Zhongyi Han, Gongxu Luo, Hao Sun, Yaqian Li, Bo Han, Mingming Gong, Kun Zhang, and Tongliang Liu. Alignclip: navigating the misalignments for robust vision-language generalization. *Machine Learning*, 114(3):58, 2025. 2

[18] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998. 6

[19] Steven A Hicks, Debesh Jha, Vajira Thambawita, Pål Halvorsen, Hugo L Hammer, and Michael A Riegler. The endotect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VIII*, pages 263–274. Springer, 2021. 6

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5

[21] Ming Hu, Jianfu Yin, Jing Wang, Yuqi Wang, Bingliang Hu, and Quan Wang. Specslice-convlstm: Medical hyperspectral image segmentation using spectral slicing and convlstm. In *International Conference on Pattern Recognition*, pages 211–225. Springer, 2024. 5

[22] Ming Hu, Jianfu Yin, Zhuangzhuang Ma, Jianheng Ma, Feiyu Zhu, Bingbing Wu, Ya Wen, Meng Wu, Cong Hu,

Bingliang Hu, et al. beta-fft: Nonlinear interpolation and differentiated training strategies for semi-supervised medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30839–30849, 2025. 1

[23] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *European Conference on Computer Vision*, pages 303–319. Springer, 2022. 1

[24] 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, 2, 3, 6, 7

[25] 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

[26] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 451–462. Springer, 2020. 6

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[28] Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. Musc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In *The Twelfth International Conference on Learning Representations*, 2024. 5

[29] Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16838–16848, 2024. 3

[30] Yunheng Li, Yuxuan Li, Quansheng Zeng, Wenhai Wang, Qibin Hou, and Ming-Ming Cheng. Densevlm: A retrieval and decoupled alignment framework for open-vocabulary dense prediction. *arXiv preprint arXiv:2412.06244*, 7, 2024. 2

[31] 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. *arXiv preprint arXiv:2301.00785*, 2023. 1

[32] Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification. *arXiv preprint arXiv:2007.01760*, 2020. 1

[33] 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. 6, 7

[34] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE, 2021. 6

[35] Shancong Mou, Xiaoyi Gu, Meng Cao, Haoping Bai, Ping Huang, Jiulong Shan, and Jianjun Shi. Rgi: robust gan-inversion for mask-free image inpainting and unsupervised pixel-wise anomaly detection. In *The Eleventh International Conference on Learning Representations*, 2022. 1

[36] Guansong Pang, Choubo Ding, Chunhua Shen, and Anton van den Hengel. Explainable deep few-shot anomaly detection with deviation networks. *arXiv preprint arXiv:2108.00462*, 2021. 1

[37] Ziyuan Qin, Huahui Yi, Qicheng Lao, and Kang Li. Medical image understanding with pretrained vision language models: A comprehensive study. *arXiv preprint arXiv:2209.15517*, 2022. 1

[38] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 30398–30408, 2025. 2

[39] 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

[40] Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2155–2162, 2023. 1

[41] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022. 1, 5

[42] 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

[43] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015. 6

[44] Alexander DJ Taylor, Phillip Tregidgo, Jonathan James Morrison, and Neill DF Campbell. Visionad, a software package of performant anomaly detection algorithms, and proportion localised, an interpretable metric. *Transactions on Machine Learning Research*, 2024. 3

[45] Yu Tian, Guansong Pang, Fengbei Liu, Yuanhong Chen, Seon Ho Shin, Johan W Verjans, Rajvinder Singh, and Gustavo Carneiro. Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 128–140. Springer, 2021. 1

[46] Yu Tian, Fengbei Liu, Guansong Pang, Yuanhong Chen, Yuyuan Liu, Johan W Verjans, Rajvinder Singh, and Gustavo Carneiro. Self-supervised pseudo multi-class pre-training for unsupervised anomaly detection and segmentation in medical images. *Medical Image Analysis*, page 102930, 2023. 1

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[48] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22883–22892, 2024. 6

[49] Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In *DAGM symposium in*, 2007. 6

[50] Guoyang Xie, Jingbao Wang, Jiaqi Liu, Feng Zheng, and Yaochu Jin. Pushing the limits of fewshot anomaly detection in industry vision: Graphcore. *arXiv preprint arXiv:2301.12082*, 2023. 1

[51] Zhen Yao, Xin Li, Taotao Jing, Shuai Zhang, and Mooi Choo Chuah. Disa: Saliency-aware foreground-background disentangled framework for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2601.20064*, 2026. 2

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

[53] Jingyi Yuan, Pengyu Jie, Junyin Zhang, Ziao Li, and Chenqiang Gao. Mfp-clip: Exploring the efficacy of multi-form prompts for zero-shot industrial anomaly detection. *arXiv preprint arXiv:2503.12910*, 2025. 5

[54] Xiaozhe Zhang, Fengying Xie, Haidong Ding, Linpeng Pan, and Zhenwei Shi. Adapt clip as aggregation instructor for image dehazing. *arXiv e-prints*, pages arXiv–2408, 2024. 5

[55] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 3

[56] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3

[57] Qihang Zhou, Shibo He, Haoyu Liu, Tao Chen, and Jiming Chen. Pull & push: Leveraging differential knowledge distillation for efficient unsupervised anomaly detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(5):2176–2189, 2023. 1

[58] 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, 2, 3, 6, 7

[59] Jiawen Zhu, Yew-Soon Ong, Chunhua Shen, and Guansong Pang. Fine-grained abnormality prompt learning for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22241–22251, 2025. 2, 6, 7

[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