

Bidirectional Multimodal Prompt Learning with Scale-Aware Training for Few-Shot Multi-Class Anomaly Detection

Yujin Lee, Sewon Kim, Daeun Moon, Seoyoon Jang, Hyunsoo Yoon*

Yonsei University, South Korea

{yjlee9040, sewon3397, dani0403, tjdb570002, hs.yoon}@yonsei.ac.kr

Abstract

Few-shot multi-class anomaly detection is crucial in real industrial settings, where only a few normal samples are available while numerous object types must be inspected. This setting is challenging as defect patterns vary widely across categories while normal samples remain scarce. Existing vision–language model–based approaches typically depend on class-specific anomaly descriptions or auxiliary modules, limiting both scalability and computational efficiency. In this work, we propose AnoPLe, a lightweight multimodal prompt learning framework that removes reliance on anomaly-type textual descriptions and avoids any external modules. AnoPLe employs bidirectional interactions between textual and visual prompts, allowing class semantics and instance-level cues to refine one another and form class-conditioned representations that capture shared normal patterns across categories. To enhance localization, we design a scale-aware prefix trained on both global and local views, enabling the prompts to capture both global context and fine-grained details. In addition, alignment loss propagates local anomaly evidence to global features, strengthening the consistency between pixel- and image-level predictions. Despite its simplicity, AnoPLe achieves strong performance on MVTec-AD, VisA, and Real-IAD under the few-shot multi-class setting, surpassing prior approaches while remaining efficient and free from expert-crafted anomaly descriptions. Moreover, AnoPLe generalizes well to unseen anomalies and extends effectively to the medical domain.

1. Introduction

In real-world industrial inspection, the cost of model development, per-category maintenance, and high-quality data collection poses significant practical challenges. Manufacturing environments often face a large number of product categories, frequent category updates, and limited opportunities to curate clean datasets. Motivated by these

*Corresponding Author.

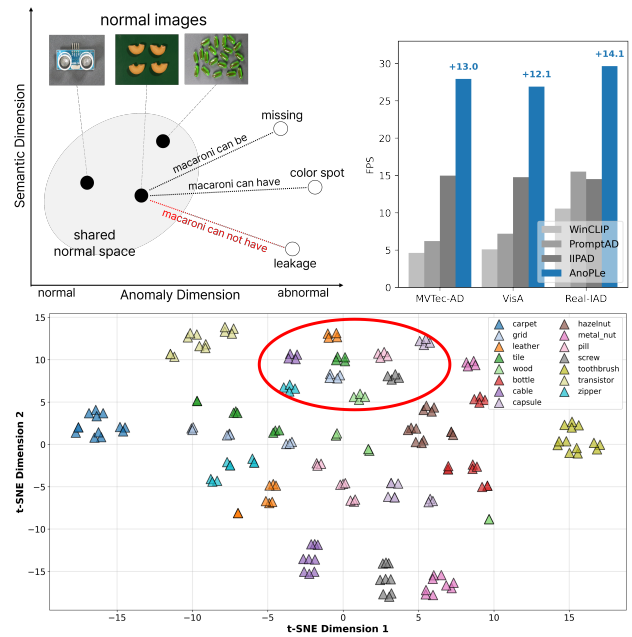


Figure 1. (a) **Category-dependent anomaly semantics (top-left).** Normal samples across categories lie in a shared normal space, while anomalies diverge along category-specific directions. Textual descriptions may align with feasible defects for each category (black), whereas invalid descriptions (red) cause semantic mismatch in text-image alignment. (b) **Inference speed comparison on MVTec-AD, VisA, and Real-IAD (top-right).** AnoPLe achieves superior efficiency by avoiding the overhead of prior methods, including WinCLIP’s sliding-window inference [17] and the additional VV-CLIP [19] or Q-Former [22] forward passes in PromptAD and IIPAD. This makes AnoPLe faster and more suitable for industrial deployment. (c) **PromptAD’s learned embeddings of MVTec-AD on multi-class setting (bottom).** T-SNE reveals that abnormal features from heterogeneous categories collapse into overlapping regions (circled), indicating semantic confusion despite distinct anomaly descriptions.

constraints, the anomaly detection (AD) community has gradually shifted from conventional single-class AD toward more realistic scenarios that better mirror industrial requirements—most notably few-shot and multi-class AD.

Few-shot anomaly detection (FSAD) aims to model normality using only a handful of verified normal samples per category, reducing the need for large curated datasets. With the emergence of vision-language models (VLMs) [18, 23], recent FSAD approaches [8, 15, 17, 19, 25] have leveraged multimodal priors through handcrafted textual templates, or learnable context vectors. While effective, these methods commonly rely on *category-specific* descriptions of defect types, such as “broken fabric” or “missing wire.” In parallel, multi-class anomaly detection (MCAD) [9, 13, 28, 30] has emerged as a scalable paradigm that unifies heterogeneous categories within a single model, reducing deployment and maintenance costs. However, it introduces a key challenge: while normal samples differ in appearance across categories, the notion of being “intact” is largely shared, whereas anomalies are diverse and category-dependent. This asymmetry makes it difficult to learn generalizable defect cues without relying on class-specific descriptions. Recent findings [7] support this view, showing that lightweight class semantics provide a strong inductive bias that organizes heterogeneous categories into a structured representation space. This suggests that category-level cues alone can effectively mitigate inter-class confusion, without handcrafted defect descriptions.

In real industrial environments, two key challenges typically arise together: only a few normal samples are available, and the system must operate across many object categories. This defines the few-shot MCAD setting [15, 22], where a single model must infer normality from limited observations and still generalize to diverse object types. While the notion of normality (free from breakage, contamination, or geometric irregularity) tends to be consistent across categories, abnormality is highly class-dependent because each category exhibits distinctive defect patterns. To illustrate the inherent limitation of description-based prompts, Figure 1-(a, c) shows how anomaly semantics vary across categories and why using generic textual descriptions often leads to mismatched text-image alignment.

Recent methods often overlook this asymmetry. IIPAD [22], for example, extends single-class AD to the multi-class scenario by collapsing category-specific anomaly descriptions into a unified pool of object-level prompts. This simplification, however, mixes heterogeneous semantics and easily produces confusion, since anomaly characteristics that should remain class-dependent are forced into a shared textual space. IIPAD introduces instance-induced prompts generated by a large Q-Former with auxiliary textual guidance, allowing prompts to adapt dynamically to each sample and partially restoring class-specific sensitivity. Although the concept of instance-induced prompts may mitigate semantic confusion, it inevitably adds substantial computational cost and still depends on descriptions derived from true abnormal samples. Thus, the limitations inherent

in description-based designs remain unresolved.

VLMs such as CLIP implicitly encode broad visual concepts of intactness and defectiveness within their latent space, acquired from massive image-text corpora. Instead of manually describing defect types or generating instance-specific prompts, a more scalable direction is to steer these latent priors through multimodal interaction. Motivated by these insights, we propose **AnoPLe**, a simple and scalable few-shot MCAD framework that avoids handcrafted anomaly descriptions and heavy auxiliary modules. We propose **bi-directional prompt learning** between textual and visual branches: class names act as lightweight semantic anchors on the textual side, while the visual branch learns instance-level cues directly from image evidence. Leveraging the reciprocal interaction, instance-level prompts refine textual cues with defect-relevant details, and textual cues guide visual features toward class-aware structure—effectively leveraging CLIP’s latent normality/abnormality priors without relying on brittle descriptions.

To enhance fine-grained localization, which is crucial in industrial inspection, AnoPLe uses **scale-aware prompt learning** that exposes the model to both global context and local details during training. Unlike VV-CLIP-based methods [19, 22, 32], our design implicitly encodes multi-scale cues into the prompt without additional inference cost. At test time, only the global prefix is used, maintaining full-image compatibility while retaining strong localization ability. Furthermore, **alignment loss** propagates pixel-level anomaly cues to the global level, promoting coherent normality modeling even under extreme few-shot constraints.

Despite requiring no expert-crafted prompts or category-specific tuning, AnoPLe achieves strong performance across diverse industrial benchmarks. Under the one-shot setting, it attains 94.5%, 86.0%, and 81.2% AUROC on MVTec-AD [3], VisA [34], and Real-IAD [26], and accurately localizes anomalies with 90.8%, 87.5%, and 88.8% PRO. The method further generalizes robustly to unseen anomaly types and even medical scenarios, suggesting applicability beyond industrial use cases. In addition, AnoPLe achieves high computational efficiency, running at 27.92, 26.90, and 29.64 FPS on MVTec-AD, VisA, and Real-IAD (Figure 1-(b)). Our main contributions are as follows:

- We propose **AnoPLe**, a unified few-shot MCAD framework that avoids handcrafted anomaly descriptions and heavy external prompt-generation modules.
- We introduce **bi-directional interactions** between textual and visual prompts to learn class-aware yet defect-agnostic representations.
- We improve localization through **scale-aware prefix** and **alignment loss** that jointly capture global context and fine-grained local cues.
- AnoPLe achieves strong few-shot performance even with the single image, showing its practicality and scalability.

2. Related work

2.1. Multi-Class Anomaly Detection

Multi-Class Anomaly Detection (MCAD) aims to unify anomaly detection across multiple object categories within a single model [9, 13, 28, 30]. Conventional one-class-one-model approaches lack scalability, motivating unified frameworks such as feature registration [15], distribution normalization [12], adaptive representation learning [28], and training-free component-aware matching frameworks [11]. A key insight from recent work is the importance of class semantics in improving discrimination across categories. CCL [7] show that explicitly modeling class-aware structure reduces inter-class confusion and yields more discriminative anomaly features, suggesting that class-level cues provide a strong inductive bias for MCAD.

Building on this, recent methods incorporate semantic guidance through vision-language models or prompt-based representations, such as AnomalyGPT and IIPAD, which leverage textual or learned semantics to better structure multi-class normality. In contrast, INP-Former [21] adopts an entirely vision-only framework, extracting intrinsic normal prototypes to guide a lightweight reconstruction pipeline. While efficient, relying solely on visual evidence limits the semantic structure available for organizing multi-class representations, which can reduce robustness to unseen categories. Together, these directions highlight two key principles for MCAD: (1) leveraging class semantics to structure multi-class representations and (2) maintaining scalability for real-world industrial settings. However, existing semantic-driven approaches often rely on explicit anomaly descriptions or additional modules, which limits their flexibility in practical few-shot settings.

2.2. Few-shot Industrial Anomaly Detection

Few-shot industrial anomaly detection aims to detect anomalies using only limited normal samples. Prior works explore transfer-based methods [15] or learning directly from limited normal samples [17, 19, 24]. With the advent of VLMs such as CLIP [18, 23], several studies adapted cross-modal alignment to improve few-shot generalization. WinCLIP [17] uses handcrafted prompts, while later works introduce learnable adapters [6, 10, 33] or context vectors [19]. These approaches demonstrate the benefit of textual conditioning but remain limited to single-category settings and often rely on expert-crafted anomaly descriptions, hindering their extension to multi-class scenarios.

Few-shot MCAD. Combining few-shot industrial anomaly detection with MCAD is challenging due to the combination of limited normal samples per category and the need to model diverse normal patterns across multiple classes within a unified representation. IIPAD [22] provides the first significant attempt in this direction by introducing

instance-induced prompts that adapt to individual samples. While effective for category-agnostic anomaly modeling, it relies on a large Q-Former and a pool of expert-defined anomaly descriptions, resulting in high computational cost and limited scalability. Building on recent MCAD findings on the importance of class-aware semantics [7], our approach leverages class names as lightweight textual anchors while avoiding anomaly-type-specific prompts. This allows AnoPLE to learn unified, class-sensitive, yet anomaly-type-agnostic representations without relying on expert-crafted anomaly descriptions or heavyweight auxiliary modules.

3. Method

AnoPLE addresses few-shot MCAD, where only a few normal samples per category are available and anomaly-specific textual descriptions are unavailable. We simulate anomalies in both pixel and latent spaces [20, 29] to learn abnormality cues from normal-only training. AnoPLE then jointly learns textual prompts for class-level semantics and visual prompts for instance-level deviations through bidirectional interactions. This design treats class names as semantic anchors while capturing shared normal patterns across categories without explicit defect descriptions. For localization, we introduce a scale-aware prefix trained on global and local crops and a lightweight decoder for pixel-level prediction, without increasing inference cost. During inference, AnoPLE further leverages a memory bank of normal features for similarity-based anomaly scoring.

3.1. Multi-Modal Prompts for Anomaly Detection

3.1.1. Textual Prompt Learning

A central challenge in few-shot MCAD is the semantic heterogeneity of abnormal patterns across categories, which renders explicit defect-level textual descriptions impractical. To establish a minimal yet effective semantic prior, we condition textual prompts solely on class names, omitting a normality token to preserve an intact class prototype. Instead, we introduce a unified abnormality token that encodes a coarse non-normal prior from CLIP pretraining, which is subsequently refined through bidirectional interactions with visual cues during forward propagation to adapt to anomaly detection. [7] show that such class semantics provide a strong inductive bias that structures categories and mitigates inter-class confusion. Moreover, class names are readily available and consistent with the anomaly detection setting, where supervision is defined by normal–abnormal distinctions rather than category identity.

Guided by these observations, we define the textual inputs for normal and abnormal prompts as

$$\mathbf{e}_0^+ = [\text{class}], \quad \mathbf{e}_0^- = [\text{abnormal}] [\text{class}], \quad (1)$$

where `class` denotes a category name and `abnormal` is a unified abnormality token shared across all categories.

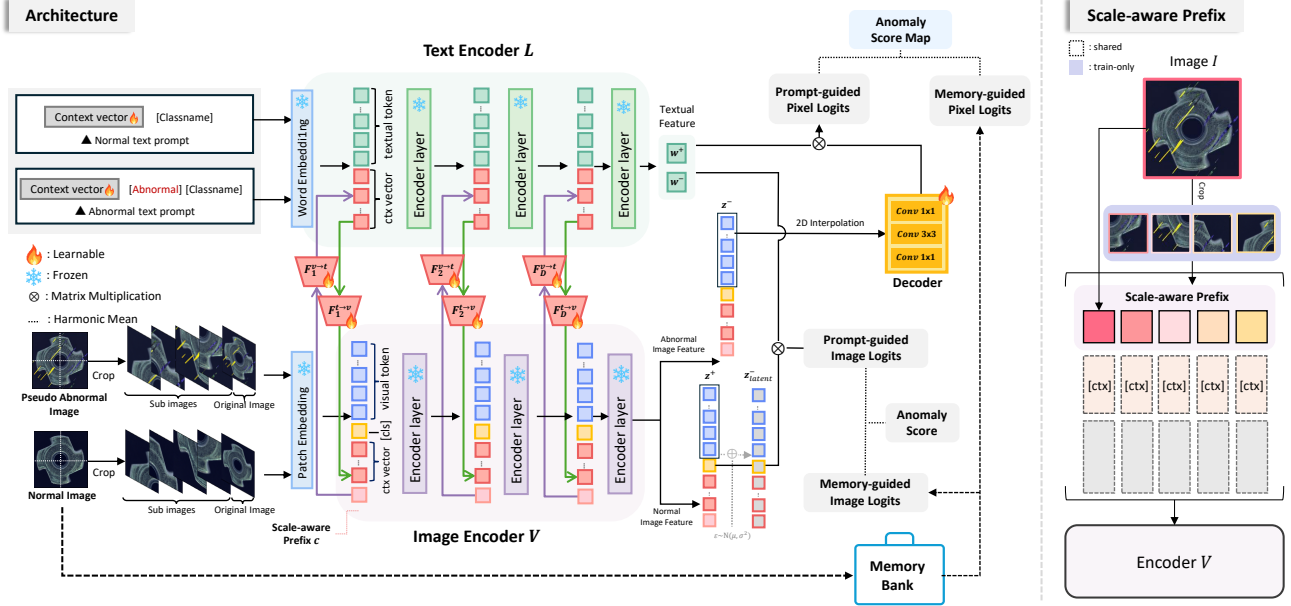


Figure 2. **Overview of AnoPLE.** AnoPLE employs bidirectional multi-modal deep prompts, establishing strong coupling between textual and visual representations learning from pseudo anomalies and unified text prompts. For localization, we introduce a light-weight decoder to produce pixel logits and train AnoPLE with mixed-scale inputs with a scale-aware prefix. At inference, the anomaly score is computed from our prompt learner and a visual memory bank. [ctx] denotes an abbreviation for the context vector.

Learnable textual context vectors $\mathbf{P}_0^t \in \mathbb{R}^{C_t \times d_t}$ are prepended to construct the initial text embeddings:

$$\mathbf{w}_0^+ = [\mathbf{P}_0^t, \mathbf{e}_0^+], \quad \mathbf{w}_0^- = [\mathbf{P}_0^t, \mathbf{e}_0^-]. \quad (2)$$

We adopt deep prompting across layers of text encoder E_T . At each layer j , the contextual vectors are concatenated with the previous hidden state:

$$\mathbf{w}_j = E_T([\mathbf{P}_j^t, \mathbf{w}_{j-1}]), \quad j = 1, \dots, J. \quad (3)$$

3.1.2. Visual Prompt Learning

Whereas textual prompts encode class-level priors, the visual branch contributes instance-level cues essential for distinguishing subtle deviations within the same category. An input image I is tokenized into patch embeddings $\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_n$, where $\mathbf{z}_0$ represents the [CLS] token capturing global image information. Learnable visual context vectors $\mathbf{P}_0^v \in \mathbb{R}^{C_v \times d_v}$ are prepended to form initial visual normal and abnormal embeddings:

$$\mathbf{V}_0^+ = [\mathbf{P}_0^v, \mathbf{z}_0^+], \quad \mathbf{V}_0^- = [\mathbf{P}_0^v, \mathbf{z}_0^-]. \quad (4)$$

Deep prompting is applied analogously to the textual branch, with E_V denoting the vision encoder:

$$\mathbf{z}_j = E_V([\mathbf{P}_j^v, \mathbf{z}_{j-1}]), \quad j = 1, \dots, J. \quad (5)$$

By operating directly at the instance level, the visual prompts capture fine-grained variations critical for modeling category-specific normality and detecting small defects.

3.1.3. Bidirectional Coupling of Multi-Modal Prompts

Since neither explicit anomaly descriptions nor true abnormal images are available, relying on a single modality may yield misaligned or incomplete representations. To enable reciprocal refinement between modalities, we design the textual and visual prompts to interact through learnable linear projections $f_{v \rightarrow t} \in \mathbb{R}^{d_v \times d_t}$ and $f_{t \rightarrow v} \in \mathbb{R}^{d_t \times d_v}$, which project visual prompts into the textual embedding space and vice versa. At each prompting layer j , we fuse the prompts from both modalities:

$$\tilde{\mathbf{P}}_j^t = [\mathbf{P}_j^t, f_{v \rightarrow t}(\mathbf{P}_j^v)], \quad \tilde{\mathbf{P}}_j^v = [\mathbf{P}_j^v, f_{t \rightarrow v}(\mathbf{P}_j^t)]. \quad (6)$$

These fused prompts are then injected into the textual and visual encoders:

$$\mathbf{w}_j = E_T([\tilde{\mathbf{P}}_j^t, \mathbf{w}_{j-1}]), \quad \mathbf{z}_j = E_V([\tilde{\mathbf{P}}_j^v, \mathbf{z}_{j-1}]). \quad (7)$$

This coupling allows class-level textual semantics and instance-level cues to iteratively guide one another, yielding representations that are both class-aware and defect-agnostic, aligning with the needs of few-shot MCAD.

3.2. Bootstrapping Local Semantics from CLIP

Scale-Aware Prompt Learning. CLIP limits its ability to capture local semantics needed for anomaly localization. Rather than modifying CLIP’s architecture or introducing

external attention modules, we incorporate scale-awareness directly into the prompt sequence. During training, we present the model with both full-resolution images I_0 and non-overlapping sub-image crops $\{I_i \mid 1 \leq i \leq N\}$ to expose it to global and local semantics. A learnable scale-aware prefix $c \in \mathbb{R}^{(N+1) \times d_v}$ encodes the resolution identity, and the corresponding c_i is prepended to the visual prompt tokens for each input. The resulting behaviors of the global and local prefixes, as well as the c , are illustrated in Fig 2. Crucially, this design maintains inference efficiency: only the global scale prefix c_{N+1} is used at test time, avoiding costly multi-scale processing. Embedding scale-awareness directly into the prompts enables CLIP to learn unified representations that respect both global and local cues in a modality-consistent and lightweight manner.

Lightweight Decoder for Local Feature Refinement. To obtain pixel-level anomaly maps from CLIP’s patch embeddings, we employ a lightweight decoder D following [31]. Patch embeddings $\mathbf{z}[1:] \in \mathbb{R}^{n \times d_v}$ are upsampled via bilinear interpolation to produce a spatial feature map $\mathbf{F} \in \mathbb{R}^{h \times w \times d_v}$, which is refined by D and projected into the text embedding dimension d_t . This refinement introduces local spatial context without increasing inference cost.

3.3. Loss Function

3.3.1. Pixel-Level Loss

Pixel-level predictions $\hat{\mathbf{M}} = \langle D(\mathbf{z}), \mathbf{w} \rangle$ are supervised using simulated ground-truth masks $\mathbf{M}$. The loss is

$$\mathcal{L}_{pixel} = \mathcal{L}_{dice}(\hat{\mathbf{M}}, \mathbf{M}) + \mathcal{L}_{focal}(\hat{\mathbf{M}}, \mathbf{M}), \quad (8)$$

where $\mathcal{L}_{dice}$ encourages overlap with anomalous regions and $\mathcal{L}_{focal}$ accounts for the rarity of anomalous pixels.

3.3.2. Image-Level Loss

We compute image-level normality scores using a contrastive cross-entropy loss:

$$\hat{p} = \frac{\exp(\langle \mathbf{z}_0^i, \mathbf{w}^i \rangle / \tau)}{\sum_{i \in \{-, +\}} \exp(\langle \mathbf{z}_0^i, \mathbf{w}^i \rangle / \tau)}, \quad (9)$$

where $i \in \{+, -\}$ denotes normal and abnormal text prompts. Text embeddings are averaged across all classes following [17]. We include pixel-space perturbations ($\mathbf{z}^-$) and latent-space perturbations ($\mathbf{z}_{latent}^- = \mathbf{z}^+ + \epsilon$ with $\epsilon \sim N(\mu, \sigma^2)$), yielding

$$\mathcal{L}_{img} = \mathcal{L}_{ce}(\mathbf{z}^+, \mathbf{z}^-, \mathbf{w}) + \mathcal{L}_{ce}(\mathbf{z}^+, \mathbf{z}_{latent}^-, \mathbf{w}). \quad (10)$$

3.3.3. Alignment Between Local and Global Semantics

To improve consistency between pixel-level and global representations, we introduce $\mathcal{L}_{align}$. Pixel logits $\hat{\mathbf{M}}$ weight the decoder outputs to compute

$$\mathbf{s} = \sum_{(i,j)} \hat{\mathbf{M}}_{ij} \circ D_{ij}(\mathbf{z}), \quad (11)$$

which aggregates local evidence aligned with anomalous regions. Alignment is enforced by cosine similarity:

$$\mathcal{L}_{align} = 1 - \langle \mathbf{z}_0, \mathbf{s} \rangle. \quad (12)$$

The final training objective is

$$\mathcal{L} = \mathcal{L}_{pixel} + \mathcal{L}_{img} + \mathcal{L}_{align}. \quad (13)$$

3.4. Anomaly Score

Following [17, 19], we compute memory-assisted anomaly maps by comparing patch features $\mathbf{F}$ from the i -th layer of E_V with a visual memory bank $\mathbf{R}$:

$$\mathbf{M}_{mem}^{ij} = \min_{\mathbf{r} \in \mathbf{R}} \frac{1}{2} (1 - \langle \mathbf{F}_{ij}, \mathbf{r} \rangle). \quad (14)$$

We fuse $\mathbf{M}_{mem}$ and the decoder output $\hat{\mathbf{M}}_q$ using a harmonic mean:

$$\mathbf{M}_q = \frac{1}{(1/\hat{\mathbf{M}}_q) + (1/\mathbf{M}_{mem})}. \quad (15)$$

The final anomaly score aggregates image-level and pixel-level signals:

$$\text{score} = \frac{1}{(1/\hat{p}) + (1/\max \mathbf{M}_q)}. \quad (16)$$

4. Experiments

4.1. Experimental Details

We conduct experiments on three widely used industrial anomaly detection benchmarks: MVTec-AD [3], VisA [34], and Real-IAD [26]. Following standard practice [3, 24], we report image-level AUROC/AUPR and pixel-level AUROC/PRO to evaluate both detection and localization performance. AnoPLe is built upon the CLIP ViT-B/16+ backbone, consistent with recent VLM-based approaches [17, 19]. For multi-scale training, input images are resized to 480×480 and split into four non-overlapping crops, while the original full-resolution view is resized to 240×240. These global and local views are used to construct the scale-aware prompting scheme. Additional hyperparameters and training details are provided in the supplementary materials.

4.2. Main Results

In Table 1, across all benchmarks, AnoPLe delivers strong performance in both image- and pixel-level metrics under the few-shot multi-class setting. Classical coreset-based methods such as SPADE [5] and PatchCore [24] degrade substantially as category diversity increases (poorly performing in Real-IAD). Recent prompt-based methods [17, 19] still suffer from semantic confusion when hand-crafted anomaly descriptions are collapsed into a unified

Table 1. **Anomaly detection and localization performance on the MVTec, VisA, and Real-IAD benchmarks.** For MVTec-AD and VisA, we reproduce each baseline under our unified training configuration. For Real-IAD, we adopt the reported values from [21]. The highest values are highlighted in bold, while the runner-up methods are underlined.

Setup	Method	MVTec				VisA				Real-IAD			
		Image-level		Pixel-level		Image-level		Pixel-level		Image-level		Pixel-level	
		AUROC	AUPR	AUROC	PRO	AUROC	AUPR	AUROC	PRO	AUROC	AUPR	AUROC	PRO
1-shot	SPADE [5]	73.5	87.5	59.2	27.3	68.2	72.3	61.5	37.6	51.2	45.6	59.5	19.3
	PatchCore [24]	66.5	80.6	88.4	66.9	69.8	70.8	88.8	70.0	59.3	55.8	89.6	60.5
	WinCLIP [17]	77.5	87.2	86.0	70.8	70.0	73.1	84.9	61.2	69.4	56.8	91.9	71.0
	PromptAD [19]	91.2	95.4	93.5	86.1	82.4	86.0	95.1	77.8	52.2	41.6	84.9	58.4
	IIPAD [22]	94.2	97.2	96.4	89.8	85.4	87.5	96.9	87.3	—	—	—	—
	INP-Former [21]	94.7	97.1	96.7	90.7	84.0	85.9	96.7	84.0	84.4	81.9	97.9	88.8
	AnoPLe	<u>94.5</u>	97.4	96.0	90.8	86.0	87.7	97.0	87.5	<u>81.2</u>	<u>80.1</u>	<u>96.7</u>	<u>88.0</u>
2-shot	SPADE [5]	75.5	88.4	59.2	27.1	69.5	74.9	61.5	37.5	50.9	45.5	59.5	19.2
	PatchCore [24]	72.7	84.8	90.6	72.5	74.0	72.7	90.1	72.1	63.3	59.7	92.0	66.1
	WinCLIP [17]	78.9	88.4	86.2	71.4	70.3	73.5	85.3	62.0	<u>70.9</u>	58.7	93.2	74.7
	PromptAD [19]	93.4	96.5	94.0	86.7	81.4	85.9	95.1	75.4	57.7	41.1	86.4	61.0
	IIPAD [22]	95.7	97.9	<u>96.7</u>	90.3	86.7	<u>88.6</u>	97.2	87.9	—	—	—	—
	INP-Former [21]	96.4	98.1	97.2	91.9	88.5	90.6	97.6	<u>87.4</u>	86.4	84.1	98.2	90.5
	AnoPLe	<u>95.8</u>	<u>98.0</u>	96.3	<u>91.4</u>	<u>86.9</u>	<u>88.6</u>	<u>97.3</u>	88.1	<u>82.8</u>	<u>82.2</u>	<u>97.4</u>	<u>89.9</u>
4-shot	SPADE [5]	77.4	89.1	59.2	27.2	70.5	73.5	61.5	37.5	50.8	45.8	59.5	19.2
	PatchCore [24]	76.8	87.8	92.1	75.8	79.3	81.8	93.0	80.0	66.0	62.2	92.9	68.6
	WinCLIP [17]	78.7	88.0	86.4	71.9	70.5	73.4	85.6	62.0	73.0	61.8	93.8	76.4
	PromptAD [19]	94.0	96.9	94.6	88.1	85.9	89.1	95.8	78.2	59.7	43.5	86.9	61.9
	IIPAD [22]	96.1	<u>98.1</u>	97.0	91.2	88.3	89.6	97.4	<u>88.3</u>	—	—	—	—
	INP-Former [21]	96.6	<u>98.1</u>	97.0	91.8	89.3	90.5	98.0	88.1	88.3	85.6	98.4	91.0
	AnoPLe	<u>96.4</u>	98.4	<u>96.5</u>	<u>91.3</u>	87.5	89.2	<u>97.5</u>	88.5	<u>83.2</u>	<u>81.5</u>	<u>97.4</u>	<u>89.5</u>

Table 2. **Results on unseen classes.** “Train” denotes the average performance on classes included during training, and “Held-out” denotes the performance on the excluded unseen class. I-AUC and P-AUC refer to Image AUROC and Pixel AUROC, respectively.

Dataset	Class	I-AUC		P-AUC	
		INP-F	AnoPLe	INP-F	AnoPLe
MVTec	Train	93.9	93.8	96.7	95.0
	Held-out	67.7 (26.2)	87.5 (6.3)	81.0 (15.7)	89.5 (5.5)
Visa	Train	87.0	83.9	97.1	95.8
	Held-out	58.7 (28.3)	72.8 (11.1)	89.7 (7.4)	91.7 (4.1)
RealIAD	Train	83.9	79.3	97.8	96.5
	Held-out	63.1 (20.8)	58.5 (20.8)	91.3 (6.5)	93.5 (3.0)

multi-class space. As particularly visualized in Figure 1-(c), PromptAD’s abnormal embeddings from different categories often collapse into overlapping clusters, indicating semantic confusion when description-based prompts are forced into a unified multi-class space. In contrast, the proposed bidirectional prompt learning enables us to extract rich multi-modal cues to be shared across categories without relying on defect-type descriptions, producing consistently robust results across datasets.

We further compare AnoPLe with INP-Former [21], a vision-only MCAD framework. Although INP-Former reports slightly higher values on a few metrics (e.g., +0.2 AUROC on MVTec image-level under 1-shot and +0.2–0.3

on VisA image-level across settings), the differences remain within a marginal range of 0.1–1.0 across all benchmarks. More importantly, its lack of textual grounding makes it less stable as category variation increases. AnoPLe instead leverages class-name prompts and multimodal interactions to maintain competitive performance across all datasets while offering significantly improved robustness to unseen categories and unseen anomaly types (Table 2). This demonstrates that multimodal prompting provides a more scalable and reliable foundation for real-world few-shot MCAD settings.

4.3. Results in Unseen Anomaly Scenarios

Real-world industrial environments frequently expose models to unseen object categories or previously unobserved anomaly types. Methods that depend on detailed defect-specific textual descriptions, such as PromptAD, tend to degrade under such conditions. AnoPLe alleviates this issue by using class names as textual anchors while avoiding dependence on anomaly-type descriptions. This design encourages the formation of category-aware yet defect-agnostic representations. In contrast, INP-Former operates purely on visual features without textual grounding, making it less flexible when encountering new categories.

To evaluate these generalization properties, we perform two complementary experiments. Table 2 presents a leave-

Table 3. **Results on unseen anomalies for selected categories from MVTec-AD and VisA, respectively.** PromptAD* excludes anomaly descriptions from the prompt pool of PromptAD.

Method	MVTec-AD		VisA	
	screw	bottle	macaroni2	pcb4
PromptAD	55.7	96.9	80.0	94.6
PromptAD*	44.7 (10.0)	93.8 (3.1)	73.7 (6.3)	90.4 (4.2)
AnoPLe (Ours)	53.9	95.1	78.5	91.2

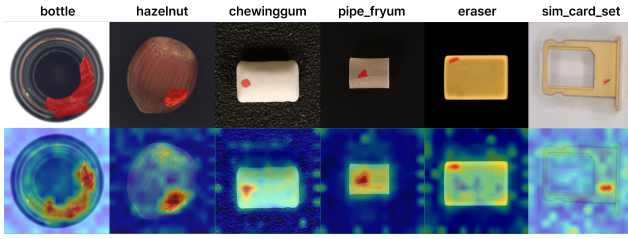


Figure 3. **Qualitative results on MVTec-AD, VisA, and Real-IAD for 1-shot pixel-level anomaly detection.** The top row shows ground-truth anomaly masks overlaid on the input images, while the bottom row displays the anomaly maps from AnoPLe.

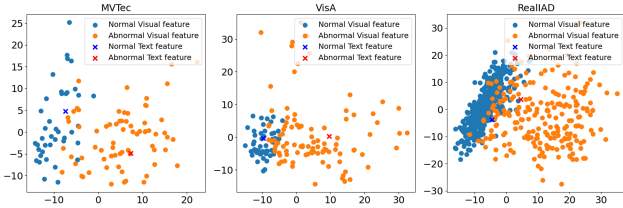


Figure 4. **Image-level Feature Visualization using PCA in a 1-shot setting.** The features “hazelnut” for MVTec, “pipe_fryum” for VisA, and “tape” for Real-IAD are used.

one-class-out analysis, where a category is excluded during training. AnoPLe shows substantially smaller performance degradation on the unseen class than INP-Former, indicating stronger category-level generalization. Table 3 examines robustness to unseen anomaly types. Removing defect-type descriptions (e.g. “manipulated front,” “with contamination,” “missing,” and “with extra”) from PromptAD to construct PromptAD* leads to large drops in I-AUROC, confirming its reliance on detailed textual priors. AnoPLe, which does not require such descriptions, remains stable even when encountering anomaly types never referenced during training. These results demonstrate that AnoPLe provides strong robustness to both unseen categories and unseen anomaly types, reinforcing its practicality for real-world industrial anomaly detection where new objects and novel defect modes routinely emerge.

4.4. Qualitative Results

As shown in Figure 3, AnoPLe successfully localizes anomalous regions even in the extremely limited 1-shot set-

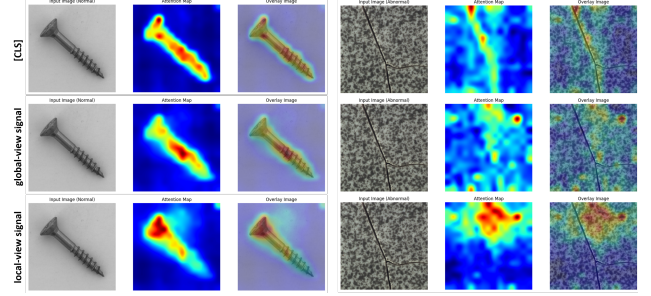


Figure 5. **Attention maps conditioned on the global-view prefix, local-view prefix, and the [CLS] token.** Columns 1, 4 are input images; columns 2, 5 display attention maps; columns 3, 6 show attention overlays.

ting, producing clean and spatially coherent heatmaps that closely align with the ground-truth regions. Despite the diverse texture, shape, and background variations across datasets, the model effectively suppresses false positives and captures fine-grained defect areas with high confidence. This demonstrates that the proposed multimodal prompt learning framework generalizes well across visual domains without category-specific tuning. Additional visual comparisons in the supplementary material show that AnoPLe provides consistently more accurate anomaly localization.

To confirm that our method performs well not only in pixel-level anomaly detection but also at the image level in a qualitative manner, Figure 4 visualizes image-level features using PCA in the 1-shot setting. The results show that AnoPLe produces well-separated clusters between normal and abnormal samples, and that textual features align naturally with their corresponding visual counterparts. This alignment confirms that the proposed bidirectional prompt learning not only enhances detection accuracy but also leads to semantically interpretable feature spaces across domains.

Beyond pixel- and image-level behaviors, we further analyze how AnoPLe utilizes multi-scale information during training. Since the model learns from both global images and local sub-images, but relies only on the global view during inference, it is essential to understand how these scale-dependent cues are organized internally. To this end, we visualize the attention maps conditioned on the learned scale-aware prefixes (Figure 5). The global-view signal attends broadly to the main object, mirroring the holistic behavior expected at test time, while the local-view signal focuses on smaller patches consistent with its restricted spatial context. Notably, the [CLS] token, trained for anomaly detection, highlights anomalous regions in defective samples while maintaining object-level focus in normal ones. These visualizations confirm that AnoPLe successfully disentangles global and local cues during training, enabling stable inference despite the inherent train–test disparity in scale.

Table 4. Ablation on context vector interaction settings, multi-scale training, scale-aware prefix c , and prompt type.

Setting	MVTec		VisA		RealID	
	I-AUC	P-AUC	I-AUC	P-AUC	I-AUC	P-AUC
(a) Interaction Settings						
Text	93.0	95.0	81.7	95.7	79.9	96.5
Image	90.3	95.2	82.0	95.8	80.3	96.5
T & I (indep)	92.4	95.2	80.3	96.0	80.1	96.6
T $\rightarrow$ I	93.5	95.1	84.2	95.9	80.1	96.5
I $\rightarrow$ T	91.2	95.0	82.0	95.9	80.0	96.5
T $\leftrightarrow$ I	94.5	96.0	86.0	97.0	81.2	96.7
(b) Alignment Strategy						
No align	93.7	95.3	86.0	95.9	73.3	94.9
align with mean	94.4	95.1	85.1	96.0	74.2	95.7
align with weighted sum	94.5	96.0	86.0	97.0	81.2	96.7
(c) Multi-scale Training / Prefix c						
single-scale & no prefix c	89.9	95.1	82.3	95.8	79.5	96.5
multi-scale & no prefix c	91.8	95.2	83.3	95.8	80.0	96.5
multi-scale & prefix c	94.5	96.0	86.0	97.0	81.2	96.7
(d) Prompt Type						
[object]	92.7	95.3	84.3	95.8	76.0	95.4
[class]	94.5	96.0	86.0	97.0	81.2	96.7

4.5. Ablations

Table 4-(a) compares interaction strategies between textual and visual context vectors. Uni-modal prompts (“Text” or “Image”) perform consistently worse, whereas cross-modal interactions bring clear improvements. The bidirectional setting (T $\leftrightarrow$ I) achieves the best results across all benchmarks (94.5 / 96.0 on MVTec-AD, 86.0 / 97.0 on VisA, 81.2 / 96.7 on Real-IAD), showing that mutual refinement between modalities yields more discriminative anomaly representations. Independent or one-way variants provide partial gains but remain less effective.

Table 4-(b) evaluates how global features should be aligned with local evidence. Skipping alignment leads to weaker normality modeling, while mean alignment partially improves results. The weighted-sum alignment performs best on all datasets by incorporating anomaly-aware pixel prototypes, providing stronger guidance for refining global semantics. Table 4-(c) shows that multi-scale training already boosts both detection and localization. Adding the scale-aware prefix c yields the highest overall performance, indicating that explicit scale encoding helps maintain consistent anomaly cues across varying object sizes.

Table 4-(d) compares class-agnostic prompts (“object”) with class-aware prompts (“class”). While the category-agnostic term appears to be more suitable for multi-class setting, class names consistently improve image-level performance (e.g., +5.2 I-AUROC on Real-IAD), as they serve as stronger semantic anchors for modeling normal object structure. We attribute this to the grounding effect of class names: they provide semantically richer anchors that help the visual prompts model normal object structure more precisely. This enhances sensitivity to subtle deviations while avoiding reliance on defect-specific anomaly descriptions, which remain harmful to generalization. Further discussion is provided in the supplementary material.

Table 5. Comparison of anomaly detection and localization performance on three medical datasets in BMAD: Brain MRI, Liver CT, and Retinal OCT (4-shots). The best score is in bold, and the second best is underlined.

Method	Brain MRI [1]		Liver CT [4, 16]		Retinal OCT [14]	
	I-AUC	P-AUC	I-AUC	P-AUC	I-AUOC	P-AUC
MedCLIP [27]	76.9	90.9	60.7	94.4	66.6	89.6
PatchCore [24]	73.4	96.3	54.9	96.0	65.3	83.2
WinCLIP [17]	53.7	86.2	61.9	97.8	75.4	94.4
PromptAD [19]	73.9	94.3	62.1	96.4	91.2	96.6
UniVAD [11]	82.6	<u>97.0</u>	<u>72.8</u>	97.9	<u>88.9</u>	<u>95.3</u>
AnoPLe	<u>79.8</u>	97.1	74.8	<u>95.9</u>	91.4	97.0

4.6. Applicability to Medical Anomaly Detection

Although AnoPLe is primarily designed for industrial domain, we demonstrate its applicability to the medical domain through the BMAD [2] under a 4-shot setting (Table 5). Despite having access to only a few normal samples per class—and without incorporating any medical-specific adaptations—AnoPLe consistently delivers strong performance across Brain MRI, Liver CT, and Retinal OCT [1, 4, 14, 16], achieving the best or second-best results. Notably, AnoPLe outperforms both medical-specialized [27] and industrial-domain-targeted anomaly detection approaches [17, 19, 24], highlighting its robustness beyond its original design scope. Moreover, AnoPLe demonstrates competitiveness comparable to the latest universal anomaly detection framework UniVAD [11], whose key strength is strong cross-domain generalization. These results indicate that AnoPLe is able to learn transferable normality cues, achieving robustness across visually diverse domains—even with limited data and no domain-specific tuning.

5. Conclusion

We present AnoPLe, a scalable framework for few-shot multi-class anomaly detection that removes the need for handcrafted anomaly descriptions through unified prompts and pseudo anomalies. By bidirectionally coupling textual and visual prompts, AnoPLe forms class-grounded and description-free representations that capture shared normal patterns across categories while detecting deviations purely from instance-level cues. To support fine-grained localization without additional inference cost, we incorporate a scale-aware prompting mechanism and a lightweight decoder that refines CLIP’s patch embeddings. Experiments on MVTec-AD, VisA, and Real-IAD show consistent gains over prior approaches in both detection and localization, along with strong robustness to unseen anomaly types, visually distinct domains, and even medical anomaly detection scenarios. By mitigating inter-class confusion and avoiding class-specific engineering, AnoPLe provides a simple, transferable, and practical solution for real-world industrial anomaly detection.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (RS-2025-02305884).

References

- [1] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 8
- [2] Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4042–4053, 2024. 8
- [3] 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. 2, 5
- [4] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical image analysis*, 84:102680, 2023. 8
- [5] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020. 5, 6
- [6] Hanqiu Deng, Zhaoxiang Zhang, Jinan Bao, and Xingyu Li. Anovl: Adapting vision-language models for unified zero-shot anomaly localization. *arXiv preprint arXiv:2308.15939*, 2023. 3
- [7] Lei Fan, Junjie Huang, Donglin Di, Anyang Su, Tianyou Song, Maurice Pagnucco, and Yang Song. Salvaging the overlooked: Leveraging class-aware contrastive learning for multi-class anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21419–21428, 2025. 2, 3
- [8] Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiguang Hu, and Jimin Xiao. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17481–17490, 2023. 2
- [9] Bin-Bin Gao. Learning to detect multi-class anomalies with just one normal image prompt. In *European Conference on Computer Vision*, pages 454–470. Springer, 2024. 2, 3
- [10] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1932–1940, 2024. 3
- [11] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Univad: A training-free unified model for few-shot visual anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15194–15203, 2025. 3, 8
- [12] Jia Guo, Haonan Han, Shuai Lu, Weihang Zhang, and Huiqi Li. Absolute-unified multi-class anomaly detection via class-agnostic distribution alignment. *arXiv preprint arXiv:2404.00724*, 2024. 3
- [13] 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. 2, 3
- [14] Junjie Hu, Yuanyuan Chen, and Zhang Yi. Automated segmentation of macular edema in oct using deep neural networks. *Medical image analysis*, 55:216–227, 2019. 8
- [15] 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. 2, 3
- [16] J Igelsias, M Styner, T Langerak, B Landman, Z Xu, and A Klein. Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015. 8
- [17] 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, 5, 6, 8
- [18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2, 3
- [19] 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. 1, 2, 3, 5, 6, 8
- [20] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411, 2023. 3
- [21] Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenying Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9974–9983, 2025. 3, 6
- [22] Wenxi Lv, Qinliang Su, and Wenchao Xu. One-for-all few-shot anomaly detection via instance-induced prompt learning. 2025. 1, 2, 3, 6
- [23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

- Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [2](#), [3](#)
- [24] 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. [3](#), [5](#), [6](#), [8](#)
- [25] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but different: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1907–1916, 2021. [2](#)
- [26] 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. [2](#), [5](#)
- [27] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, 2022. [8](#)
- [28] 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](#), [3](#)
- [29] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem—a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021. [3](#)
- [30] Ying Zhao. Omnia: 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](#), [3](#)
- [31] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. [5](#)
- [32] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*, 2024. [2](#)
- [33] Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17826–17836, 2024. [3](#)
- [34] 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. [2](#), [5](#)