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.

VisualAD: Language-Free Zero-Shot Anomaly Detection via Vision Transformer

Yanning Hou[3,2] Peiyuan Li[2] Zirui Liu[2] Yitong Wang[2]
Yanran Ruan[2] Jianfeng Qiu[2] Ke Xu[1,2*]

[1] State Key Laboratory of Opto-Electronic Information Acquisition and Protection Technology, Anhui University, Hefei, China
[2] School of Artificial Intelligence, Anhui University, Hefei, China
[3] College of Intelligence Science and Technology, National University of Defense Technology, Changsha, China

Abstract

Zero-shot anomaly detection (ZSAD) requires detecting and localizing anomalies without access to target-class anomaly samples. Mainstream methods rely on vision–language models (VLMs) such as CLIP: they build hand-crafted or learned prompt sets for normal and abnormal semantics, then compute image–text similarities for open-set discrimination. While effective, this paradigm depends on a text encoder and cross-modal alignment, which can lead to training instability and parameter redundancy. This work revisits the necessity of the text branch in ZSAD and presents VisualAD, a purely visual framework built on Vision Transformers. We introduce two learnable tokens within a frozen backbone to directly encode normality and abnormality. Through multi-layer self-attention, these tokens interact with patch tokens, gradually acquiring high-level notions of normality and anomaly while guiding patches to highlight anomaly-related cues. Additionally, we incorporate a Spatial-Aware Cross-Attention (SCA) module and a lightweight Self-Alignment Function (SAF): SCA injects fine-grained spatial information into the tokens, and SAF recalibrates patch features before anomaly scoring. VisualAD achieves state-of-the-art performance on 13 zero-shot anomaly detection benchmarks spanning industrial and medical domains, and adapts seamlessly to pretrained vision backbones such as the CLIP image encoder and DINOv2. Our code is publicly available at https://github.com/7HHHHH/VisualAD.

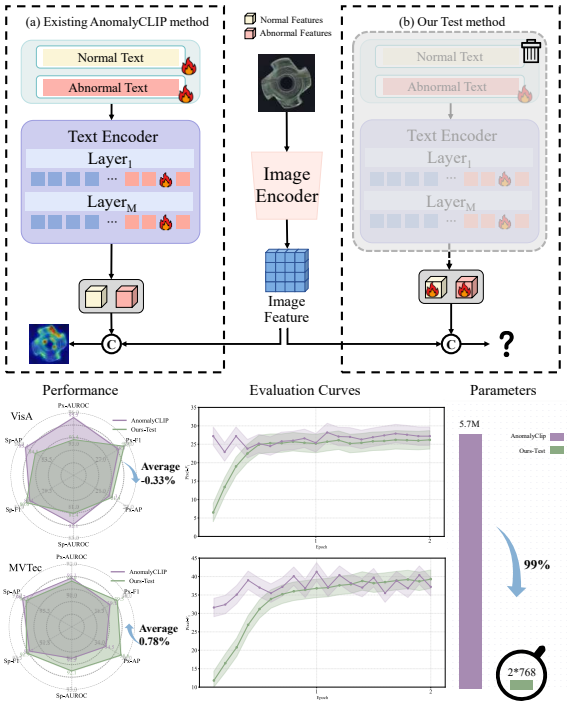


Figure 1. Exploratory study motivating this work. (a) AnomalyCLIP derives normal/abnormal prototypes via trainable text prompts and a text encoder. (b) A purely visual variant removes the text branch and directly learns two visual prototypes, achieving comparable or slightly better results on VisA and MVTec with 99% fewer parameters. Radar and line plots show that our variant maintains similar accuracy but with much smoother evaluation curves, whereas AnomalyCLIP oscillates noticeably.

1. Introduction

Anomaly detection [38, 45, 50, 52] in images plays a central role in safety-critical domains such as industrial inspection and medical diagnosis. The objective is to identify deviations from expected patterns at either the image or pixel level, from subtle surface defects on manufactured parts

* Corresponding author.

to early pathological changes in medical scans. Most established methods follow unsupervised [19, 35, 48, 56] or semi-supervised [32, 59] paradigms that assume access to category-specific normal images, labeled anomalies, or both during training. In realistic deployments, however, a cold-start situation is common: collecting sufficient, clean normal data for every new product line or disease category is costly and sometimes impractical. As a result, these methods may be difficult to scale to long-tail categories and are often brittle under domain shifts between carefully curated training data and noisy real-world environments. Zero-shot anomaly detection (ZSAD) [9, 15, 17, 58] seeks to alleviate these limitations by detecting anomalies in previously unseen categories without using any in-category images for training, thereby decoupling anomaly detection from the per-category data collection burden.

Against this backdrop, large-scale pretrained models [10, 39] offer a practical route to ZSAD. Pretrained vision–language models (VLMs) [29, 30, 43] such as CLIP [43] exhibit strong transferability by aligning visual and textual semantics in a shared embedding space. Leveraging this property, a line of work introduces prompts denoting normality and abnormality into the text encoder and contrasts the resulting text embeddings with image features to perform open-set anomaly detection. Within this paradigm, textual prompts are typically divided into two categories. The hand-crafted [11, 12, 22] category, exemplified by WinCLIP [24], relies on fixed templates that combine category context words with state descriptors to express normal and abnormal states; although training-free and easy to apply, such designs are sensitive to wording choices and may struggle to cover diverse anomaly patterns. The learnable category [9, 17, 58], exemplified by AnomalyCLIP [58], parameterizes contextual terms as trainable vectors and optimizes them on a small auxiliary dataset, yielding object-agnostic representations of normality and abnormality that can generalize across categories. Despite their differences, these methods suggest a simple principle: anomalies can be described by two sets of reference features corresponding to normality and abnormality.

However, this principle raises a question: if the final decision is governed by only two sets of latent vectors, normal and anomaly, is the language modality truly indispensable? From a purely visual standpoint, anomalies appear as structural or statistical deviations in texture, shape, or color, all of which can be captured within the visual domain. To probe this hypothesis, we conducted a small-scale study by modifying AnomalyCLIP [58]: we removed the text encoder and directly optimized two learnable vectors representing normality and abnormality. The resulting detection performance showed negligible degradation, while the number of trainable parameters dropped by more than 99%. As illustrated in Fig. 1, the evaluation curves across training epochs reveal a clear difference in generalization behavior: our visual-only variant improves steadily and maintains a smooth upward trend, whereas the original AnomalyCLIP displays pronounced fluctuations across epochs. This finding suggests that, in CLIP-based ZSAD pipelines, textual prompts might serve primarily as an indirect route for shaping a pair of discriminative visual prototypes, rather than providing essential semantic grounding.

Motivated by this observation, we propose VisualAD, a purely visual framework for zero-shot anomaly detection built on frozen Vision Transformers. Instead of relying on a text encoder, VisualAD introduces two learnable tokens, an anomaly token and a normal token, directly into the ViT token sequence. Through multi-layer self-attention, these tokens interact with patch tokens, gradually acquiring high-level notions of normality and anomaly while guiding patches to emphasize anomaly-related cues. To align the high-level semantic tokens with fine-grained image features across layers, we employ a Spatial-Aware Cross-Attention (SCA) module that provides explicit spatial cues and local detail to the tokens, together with a lightweight Self-Alignment Function (SAF), implemented as a small MLP, that recalibrates patch features before token–patch alignment. Unlike methods that align directly to patch-level features, SCA dynamically updates the tokens using fine-grained visual evidence, which improves ZSAD performance. We then synthesize the anomaly map by integrating alignment outcomes between multiple updated tokens and recalibrated patch-level features across layers, and obtain the image-level anomaly decision from the top one percent of pixels with the highest anomaly scores.

In summary, our contributions are as follows:

- We revisit the necessity of text in zero-shot anomaly detection and show that discriminative anomaly features can be learned purely from visual cues.
- We introduce VisualAD, a ViT-only framework for zero-shot anomaly detection that injects two learnable tokens into a frozen backbone, allowing them to interact with patch tokens through multi-layer self-attention to encode normality and abnormality.
- We propose the SCA and SAF modules, where SCA injects explicit spatial evidence into the tokens and SAF recalibrates patch features, enabling stable multi-layer alignment and improved localization.
- Extensive experiments across industrial and medical benchmarks demonstrate strong zero-shot generalization to unseen categories and datasets.

2. Related work

2.1. Unsupervised Anomaly Detection.

Unsupervised anomaly detection seeks to identify anomalies using only normal training images. Existing meth-

ods fall into three categories: embedding-based, discriminative, and reconstruction-based approaches. Embedding-based methods, such as PaDiM [13], PatchCore [44], and CSFlow [46], utilize features from pre-trained models to distinguish anomalies from normal samples. Discriminative methods, like CutPaste [28], introduce synthetic anomalies to convert the task into a supervised problem. Reconstruction-based [16, 54, 55] methods, including autoencoders [5, 18, 23] and GANs [40, 53, 57], assume anomalies yield higher reconstruction errors due to their absence in normal data. Recent knowledge distillation and feature reconstruction approaches balance efficiency and performance by aligning a student network with a pre-trained teacher. However, these methods struggle with unseen anomaly types, requiring extensive normal data collection and retraining for new scenarios, which limits their practical adaptability.

2.2. Zero-shot Anomaly Detection.

In industrial anomaly detection, pre-trained vision models [10, 14, 27, 29, 31, 43] show strong generalization and feature extraction. Zero-shot approaches fall into two families. The first is training-free: WinCLIP [24] aligns multi-granularity features but needs repeated encodings; SAA [8] couples Grounding DINO [33] with SAM [27] for precise localization and segmentation at higher computational cost. The second family keeps the target domain zero-shot while training lightweight modules on auxiliary data to align vision and text. AnomalyCLIP [58] uses object-agnostic prompts with glocal optimization but relies on a dual branch; AdaCLIP [9] and VCP-CLIP [41] inject visual cues into text prompts, improving adaptability with added complexity; CLIP-AD [12] and CLIP-SAM [21] further tighten alignment or add segmentation priors. AA-CLIP [34] adopts a two-stage scheme with anomaly and normal text anchors plus lightweight visual adapters; AdaptCLIP [17] learns compact visual, textual, and prompt–query adapters for cross-domain transfer; Bayes-PFL [42] models a distribution over prompts to capture diverse anomaly semantics. In contrast, VisualAD removes the text branch and learns high-level anomaly cues purely from vision, achieving strong zero-shot results.

3. Approach

3.1. Problem Setting

We follow the zero-shot anomaly detection (ZSAD) protocol. The model is trained with supervision on visible categories C_s drawn from industrial datasets and is directly evaluated on unseen categories C_u collected from other industrial or medical datasets. By definition $C_u \cap C_s = \varnothing$; the auxiliary training and testing categories come from distinct datasets that exhibit substantial domain shifts. This

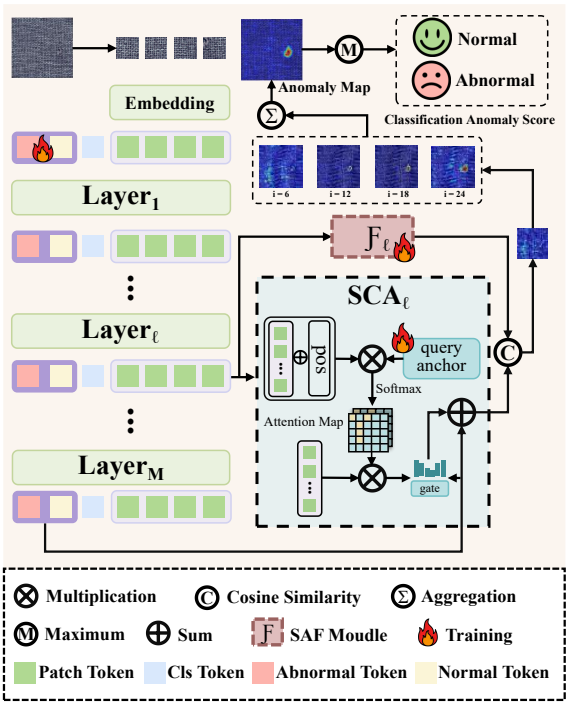


Figure 2. **Overview of VisualAD.** Two learnable global tokens (anomaly and normal) are inserted into a frozen ViT. For intermediate layers ℓ, SCA uses a few anchor queries with positional encoding and token-guided gating to aggregate localized spatial evidence, yielding enhanced tokens $\tilde{\mathbf{t}}_a^{(\ell)}, \tilde{\mathbf{t}}_n^{(\ell)}$. In parallel, each layer's patch features are recalibrated by a SAF, denoted as $\mathcal{F}_\ell$. The cosine-similarity difference between enhanced tokens and recalibrated patches forms layer-wise anomaly maps, which are upsampled and summed to obtain the final anomaly map; the image-level score is the mean of the top-k responses.

setting evaluates cross-domain generalization without any category-specific fine-tuning.

As illustrated in Fig. 2, we propose **VisualAD**, a zero-shot anomaly detection framework built upon frozen Vision Transformers. Given a backbone such as ViT-L/14@336px, we insert two learnable global tokens, the anomaly token $\mathbf{t}_a$ and the normal token $\mathbf{t}_n$, to directly encode abnormality and normality in the visual feature space. The input sequence is

$$\mathbf{z}_0 = [\mathbf{t}_a,\, \mathbf{t}_n,\, \mathbf{t}_c,\, \mathbf{p}_1,\, \ldots,\, \mathbf{p}_N], \tag{1}$$

where $\mathbf{t}_c$ is the class token, $\{\mathbf{p}_i\}_{i=1}^N$ are image patch tokens, and all tokens are d-dimensional.

To capture diverse spatial and semantic cues, we extract features from a set of intermediate layers $L = \{l_1, \ldots, l_K\}$ (default $\{6, 12, 18, 24\}$ for ViT-L/14). For each selected layer, we introduce a *Spatial-Aware Cross-Attention* (SCA) module that aggregates localized spatial evidence from patch features and enhances the token representations. In parallel, a lightweight *Self-Alignment Function* (SAF) re-

calibrates patch features at layer ℓ. The enhanced tokens and the SAF-recalibrated patches jointly produce layer-wise anomaly maps that are fused across layers. Pixel-level and image-level anomaly scores are obtained from the fused map. Training and inference share the same computation path with no text encoder or cross-modal alignment.

3.2. Spatial-Aware Cross-Attention

Global tokens encode high-level semantics but lack explicit spatial grounding. To inject localized evidence while keeping the ViT pipeline intact, we apply a single SCA module at each selected layer ℓ. Given patch features $\mathbf{P}_\ell \in \mathbb{R}^{B\times N\times d}$, where B is the batch size, N is the number of patches, and d is the feature dimension, we first augment them with a layer-specific learnable positional encoding $\mathbf{E}^{(\ell)}_{\text{pos}}$:

$$\mathbf{P}^{\text{pos}}_\ell = \mathbf{P}_\ell + \mathbf{E}^{(\ell)}_{\text{pos}}. \tag{2}$$

allowing the module to distinguish spatially distinct regions.

To aggregate spatial cues efficiently, SCA employs m learnable anchor queries $\mathbf{Q}_{\text{anchor}} \in \mathbb{R}^{m\times d}$ with $m \ll N$. For clarity, we omit the batch dimension in the following and denote $\mathbf{P}^{\text{pos}}_\ell \in \mathbb{R}^{N\times d}$. The cross-attention weights and anchor-aggregated features are computed as

$$\mathbf{A}_\ell = \text{softmax}\left(\frac{\mathbf{Q}_{\text{anchor}}(\mathbf{P}^{\text{pos}}_\ell)^\top}{\sqrt{d}}\right), \qquad \mathbf{U}_\ell = \mathbf{A}_\ell\,\mathbf{P}_\ell. \tag{3}$$

where $\mathbf{A}_\ell \in \mathbb{R}^{m\times N}$ is the spatial attention distribution and $\mathbf{U}_\ell \in \mathbb{R}^{m\times d}$ contains the anchor-aggregated features, whose i-th row is denoted by $\mathbf{a}_i$.

Next, SCA adapts the anchor features to each token via a token-guided gating mechanism. We compute a gate vector

$$\mathbf{g}(\mathbf{t}) = \sigma(W_g\mathbf{t}) \in \mathbb{R}^m, \tag{4}$$

where $\mathbf{t} \in \{\mathbf{t}_a, \mathbf{t}_n\}$ is a global token before enhancement, and W_g is shared across anchors. The enhanced token is

$$\tilde{\mathbf{t}}_\ell = \mathbf{t} + \alpha\sum_{i=1}^{m} g_i(\mathbf{t})\,\mathbf{a}_i, \tag{5}$$

with α a learnable residual scale and $g_i(\mathbf{t})$ the i-th entry of $\mathbf{g}(\mathbf{t})$. This formulation preserves the semantic role of global tokens while selectively injecting spatial information in an anchor-specific manner.

We instantiate SCA independently for each $\ell \in \mathcal{L}$, producing $\tilde{\mathbf{t}}^{(\ell)}_a$ and $\tilde{\mathbf{t}}^{(\ell)}_n$ that are subsequently matched with the recalibrated patch features to form the layer-wise anomaly maps. Since attention and gating are recomputed for every image, SCA dynamically adapts the abnormality sensitivity of the global tokens to the local structure of each test sample.

3.3. Self-Alignment Function and Anomaly Scoring

Beyond SCA, we refine patch representations using a learnable *Self-Alignment Function (SAF)*. For each selected layer ℓ, the SAF is implemented as a one-hidden-layer MLP with hidden dimension equal to the input size:

$$\hat{\mathbf{P}}_\ell = \mathcal{F}_\ell(\mathbf{P}_\ell), \tag{6}$$

where $\mathcal{F}_\ell$ denotes the SAF at layer ℓ. This nonlinear recalibration aligns patch features with the evolving normal and anomaly tokens.

Patch-level anomaly scores are then computed through a self-aligned cosine contrast. Using ℓ_2-normalized features ($\bar{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\|_2$), the score of the i-th patch at layer ℓ is

$$s^{(\ell)}_i = \langle \bar{\hat{\mathbf{p}}}^{(\ell)}_i,\, \bar{\mathbf{t}}^{(\ell)}_a \rangle - \langle \bar{\hat{\mathbf{p}}}^{(\ell)}_i,\, \bar{\mathbf{t}}^{(\ell)}_n \rangle. \tag{7}$$

Scores are reshaped into spatial maps $\mathbf{H}_\ell \in \mathbb{R}^{H\times W}$ and upsampled to the input size. Multi-layer fusion yields the final anomaly map:

$$\mathbf{H} = \sum_{\ell\in\mathcal{L}} \mathbf{H}_\ell. \tag{8}$$

The image-level anomaly score S is computed by averaging the top-k most anomalous pixels, where $k = \lfloor 0.01\,HW \rceil$ denotes the floor of 1% of all pixels:

$$S = \frac{1}{k}\sum_{i=1}^{k} \mathbf{H}_{(i)}, \tag{9}$$

with $\mathbf{H}_{(i)}$ the i-th largest value in $\mathbf{H}$. This completes the construction of both the anomaly map and the corresponding image-level score in VisualAD.

3.4. Training Objective

We jointly optimize image-level classification, pixel-level segmentation, and token contrastive separation under a unified objective, while keeping the ViT backbone frozen; only the anomaly/normal tokens $\mathbf{t}_a, \mathbf{t}_n$, the SCA modules, and the per-layer transformations $\{\mathcal{F}_\ell\}$ are updated. For image-level supervision, we apply binary cross-entropy to the score S:

$$\mathcal{L}_{\text{cls}} = -\Big[y\log\sigma(S) + (1-y)\log\big(1-\sigma(S)\big)\Big]. \tag{10}$$

For pixel-level supervision, we combine Focal Loss and Dice Loss on each selected layer map, where $\hat{\mathbf{M}}^{(\ell)} = \sigma(\mathbf{H}_\ell)$ denotes the predicted mask and $\mathbf{M}$ is the binary ground truth:

$$\mathcal{L}_{\text{seg}} = \sum_{\ell\in\mathcal{L}} \left(\mathcal{L}^{(\ell)}_{\text{focal}} + \mathcal{L}^{(\ell)}_{\text{dice}}\right). \tag{11}$$

To explicitly separate the anomaly and normal tokens, we impose a cosine-margin penalty at the deepest layer index L (the maximum element in $\mathcal{L}$):

$$\mathcal{L}_{\text{ctr}} = \max\left(0,\ \langle \bar{\mathbf{t}}_a^{(L)}, \bar{\mathbf{t}}_n^{(L)} \rangle + \tau\right), \quad \tau = 0.5, \tag{12}$$

which encourages their similarity to fall below -0.5, corresponding to an angular distance greater than $120°$. The overall loss is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{ctr}}. \tag{13}$$

Training and inference strictly share the same pipeline, including the layer set $\mathcal{L}$, SCA, the transformations $\mathcal{F}_\ell$, cross-layer fusion, and top-k aggregation, with no text branch or cross-modal alignment.

4. Experiments

4.1. Experimental Setup

Datasets. To evaluate the zero-shot anomaly detection (ZSAD) performance of our model, we conduct experiments on 13 real-world datasets spanning both industrial and medical domains. In the industrial domain, we employ six widely used benchmarks: MVTec-AD [6], VisA [60], BTAD [37], KSDD2 [49], DAGM [51], and DTD-Synthetic [1]. For the medical domain, we adopt OCT17 [26], BrainMRI [47], BrainAD [2, 3, 36], HIS [4], CVC-ClinicDB [7], Endo [20], and Kvasir [25], covering diverse imaging modalities and tasks. Specifically, OCT17 contains retinal OCT scans for retinal disease classification, BrainMRI and BrainAD provide brain MRI images for brain tumor or lesion classification, and HIS consists of histopathology images for abnormal tissue analysis. CVC-ClinicDB, Endo, and Kvasir are colonoscopy datasets with pixel-level annotations for colon polyp segmentation. Following previous works, we train the model on one industrial dataset and directly perform inference on other industrial and medical datasets without further fine-tuning. Since VisA contains object categories disjoint from the others, it is used as the auxiliary training set. For evaluating VisA itself, the model is fine-tuned on MVTec-AD. Detailed descriptions of all datasets are provided in Appendix.

Evaluation metrics. For classification, we report image-level metrics including the Area Under the ROC Curve (AUROC), the maximum F_1 score over all thresholds (F_1-max), and Average Precision (AP). For segmentation, we evaluate pixel-level AUROC, the maximum F_1 score over all thresholds (F_1-max), AP, and Per-Region Overlap (PRO) to assess the localization capability of the model.

Implementation Details. We adopt publicly available backbones: CLIP (ViT-L/14@336px) and DINOv2 (ViT-L/14). Following standard practice, input images are resized to 518×518. From the 24-layer visual encoder, we extract patch tokens at layers $\{6, 12, 18, 24\}$. The number of anchor queries in SCA is set to $m = 4$ by default. VisualAD is trained with the Adam optimizer, using an initial learning rate of 1×10^{-3} and a batch size of 8. Additional implementation details are provided in Appendix.

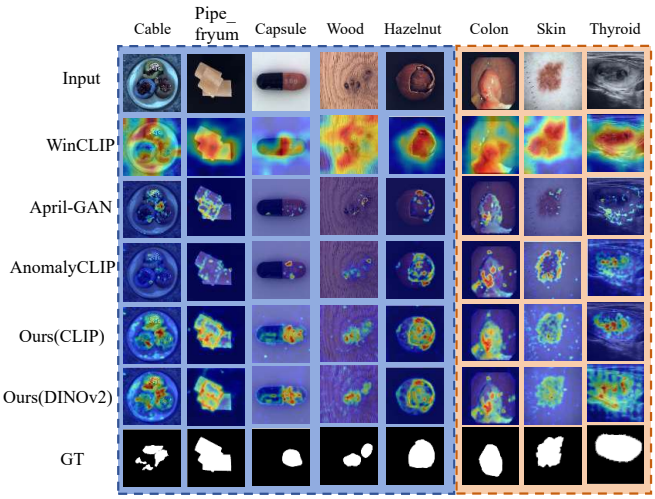


Figure 3. Qualitative comparison of anomaly segmentation results for different ZSAD methods. The first five columns show images from industrial datasets, while the last three columns correspond to medical datasets.

4.2. Performance Comparison with SOTA Method

In this study, we compare VisualAD with five state-of-the-art methods: WinCLIP [24], APRILGAN [11], CLIP-AD [12], AnomalyCLIP [58], and AdaCLIP [9]. Since our approach can adapt to multiple backbones, we report results using both CLIP's ViT-L/14@336px and DINOv2 (ViT-L/14) to demonstrate its flexibility across different visual architectures. Table 1 reports quantitative ZSAD results on industrial and medical benchmarks. Compared with existing methods, **VisualAD** achieves state-of-the-art performance on almost all datasets at both the image and pixel levels. On industrial datasets in particular, VisualAD with CLIP ViT-L/14@336px attains the best results across all classification metrics. We further observe that CLIP-based and DINOv2-based instances of VisualAD exhibit complementary strengths, with the former slightly favoring image-level classification and the latter delivering stronger pixel-wise segmentation performance. Fig. 3 visualizes anomaly maps from representative industrial and medical datasets. Compared with other approaches, VisualAD produces more accurate segmentations and more complete localization of anomalous regions. The advantage is especially pronounced on medical categories (for ex-

Table 1. Comparison of zero-shot anomaly detection methods on industrial and medical benchmarks. For each dataset, we report image-level metrics (AUROC, F_1-max, AP) and pixel-level metrics (AUROC, F_1-max, AP, PRO). VisualAD(CLIP) and VisualAD(DINOv2) denote our method instantiated with CLIP and DINOv2 backbones, respectively. The best and second-best results for each metric and dataset are highlighted in red and blue.

Domain	Metric	Dataset	WinCLIP [24]	APRIL-GAN [11]	CLIP-AD [12]	AnomalyCLIP [58]	AdaCLIP [9]	VisualAD(CLIP [43])	VisualAD(DINOv2 [39])
			ViT-L/14@336px	ViT-L/14@336px	ViT-L/14@336px	ViT-L/14@336px	ViT-L/14@336px	ViT-L/14@336px	ViT-L/14
Industrial	Image-level (AUROC, F_1-max, AP)	MVTec-AD [6]	(90.4, 92.7, 95.6)	(86.1, 90.4, 93.6)	(74.1, 86.3, 88.1)	(91.6, 92.7, 96.2)	(92.0, 92.7, 96.4)	(92.2, 93.2, 96.7)	(90.1, 92.4, 94.8)
		VisA [60]	(75.6, 78.2, 78.8)	(77.4, 78.6, 80.9)	(66.2, 74.3, 71.4)	(81.0, 80.3, 84.4)	(79.7, 79.63, 83.2)	(84.7, 82.5, 87.6)	(83.1, 81.4, 86.8)
		BTAD [37]	(68.2, 67.8, 70.9)	(73.7, 68.7, 69.9)	(66.7, 65.9, 67.3)	(88.7, 86.0, 90.6)	(90.0, 87.2, 91.5)	(94.9, 93.9, 97.0)	(88.2, 84.7, 89.7)
		KSDD2 [49]	(93.5, 86.4, 94.2)	(90.4, 82.9, 92.0)	(81.7, 75.0, 85.5)	(91.9, 84.5, 93.4)	(94.9, 90.3, 96.2)	(98.0, 93.9, 98.3)	(97.7, 93.1, 98.1)
		DAGM [51]	(91.8, 75.8, 79.5)	(94.4, 80.3, 83.9)	(62.1, 37.1, 32.3)	(98.0, 90.6, 92.4)	(98.3, 91.5, 94.2)	(99.5, 95.0, 97.8)	(93.2, 83.9, 86.1)
		DTD-Synthetic [1]	(95.1, 94.1, 97.7)	(85.5, 89.1, 94.0)	(75.1, 86.3, 88.0)	(93.7, 94.3, 97.4)	(92.1, 92.4, 96.3)	(97.5, 96.6, 99.1)	(91.0, 94.4, 97.4)
	Pixel-level (AUROC, F_1-max, AP, PRO)	MVTec-AD [6]	(82.3, 24.8, 18.2, 62.0)	(87.5, 42.3, 39.1, 43.7)	(77.9, 26.3, 21.1, 55.7)	(91.0, 38.9, 34.4, 81.7)	(88.5, 43.9, 41.0, 47.6)	(90.8, 43.9, 41.2, 87.5)	(91.3, 47.4, 45.4, 88.6)
		VisA [60]	(73.2, 9.0, 5.4, 51.1)	(93.8, 32.6, 26.2, 86.5)	(93.0, 24.1, 17.9, 80.2)	(95.4, 27.6, 20.7, 86.4)	(95.1, 33.8, 29.2, 71.3)	(95.8, 34.6, 28.4, 91.0)	(95.3, 35.2, 29.9, 88.2)
		BTAD [37]	(72.7, 18.5, 12.9, 27.3)	(91.3, 40.1, 37.7, 21.0)	(80.9, 24.1, 18.3, 41.4)	(93.0, 47.1, 41.5, 71.0)	(87.7, 42.3, 36.6, 17.1)	(91.1, 49.8, 43.1, 80.4)	(93.4, 42.6, 38.7, 76.7)
		KSDD2 [49]	(94.1, 24.6, 17.4, 77.6)	(94.5, 64.2, 66.9, 39.2)	(95.6, 43.0, 39.6, 73.5)	(98.0, 50.6, 43.7, 90.8)	(96.1, 59.2, 58.6, 40.8)	(98.5, 60.9, 62.1, 98.5)	(98.9, 64.0, 66.8, 98.9)
		DAGM [51]	(87.6, 12.7, 6.8, 65.7)	(84.4, 35.1, 27.8, 12.7)	(69.1, 20.9, 14.7, 36.1)	(96.9, 56.9, 53.6, 89.2)	(88.6, 48.5, 42.6, 37.6)	(92.2, 57.9, 56.4, 89.3)	(89.5, 54.8, 52.2, 84.5)
		DTD-Synthetic [1]	(79.5, 16.1, 9.8, 51.5)	(94.9, 60.4, 61.0, 33.8)	(86.6, 35.8, 31.0, 63.2)	(97.5, 55.8, 52.5, 87.9)	(95.1, 58.4, 56.1, 34.3)	(98.1, 64.3, 65.5, 94.8)	(96.7, 65.8, 67.7, 92.4)
Medical	Image-level (AUROC, F_1-max, AP)	OCT17 [26]	(55.2, 82.7, 81.1)	(30.4, 81.7, 67.0)	(58.1, 83.6, 84.1)	(63.7, 85.7, 86.5)	(77.3, 86.1, 91.8)	(88.9, 88.2, 96.4)	(91.2, 90.3, 97.1)
		BrainMR1 [47]	(86.6, 84.1, 91.5)	(89.3, 88.2, 90.9)	(83.0, 81.9, 89.3)	(96.4, 94.7, 97.2)	(94.9, 93.1, 96.8)	(96.7, 94.8, 97.6)	(93.8, 90.3, 95.9)
		Brain_AD [2]	(72.5, 90.7, 91.6)	(58.8, 90.6, 87.8)	(72.1, 91.3, 91.5)	(69.0, 90.6, 90.1)	(80.0, 90.2, 94.1)	(80.8, 91.8, 94.7)	(87.1, 92.5, 96.7)
		HIS [4]	(47.1, 45.6, 40.9)	(59.8, 67.5, 56.5)	(44.7, 66.6, 46.0)	(55.2, 66.8, 56.1)	(59.9, 68.4, 55.8)	(60.1, 68.9, 55.9)	(60.1, 68.7, 56.2)
	Pixel-level (AUROC, F_1-max, AP, PRO)	Brain_AD [2]	(87.6, 21.7, 13.3, 59.7)	(83.6, 38.7, 35.0, 45.4)	(94.1, 42.9, 40.5, 74.2)	(95.1, 43.1, 42.3, 71.5)	(95.2, 40.6, 37.0, 36.4)	(95.2, 46.7, 43.7, 79.5)	(96.4, 50.2, 51.9, 83.5)
		CVC-ClinicDB [7]	(70.3, 27.3, 19.4, 32.5)	(82.4, 38.2, 36.4, 53.5)	(76.5, 32.1, 24.6, 52.8)	(84.6, 45.5, 41.7, 69.8)	(84.3, 44.4, 43.3, 63.6)	(85.2, 45.6, 37.8, 70.6)	(85.9, 45.0, 36.2, 67.3)
		Endo [20]	(68.2, 32.9, 23.8, 28.3)	(82.7, 46.9, 47.2, 48.3)	(78.1, 41.4, 36.8, 46.4)	(86.5, 53.7, 50.7, 67.1)	(82.9, 47.9, 47.7, 55.8)	(84.9, 52.0, 46.2, 66.3)	(86.8, 54.7, 53.2, 67.8)
		Kvasir [25]	(69.7, 35.9, 27.8, 24.5)	(77.6, 43.2, 42.5, 40.6)	(73.6, 38.9, 31.1, 30.6)	(82.0, 50.0, 43.2, 44.7)	(80.3, 47.7, 31.0, 27.5)	(80.3, 47.8, 47.6, 50.8)	(82.6, 50.3, 46.4, 40.5)

ample, skin), where our method yields clearer boundaries and fewer false positives. These results indicate that high-level anomaly semantics learned purely in the visual feature space can generalize effectively across domains.

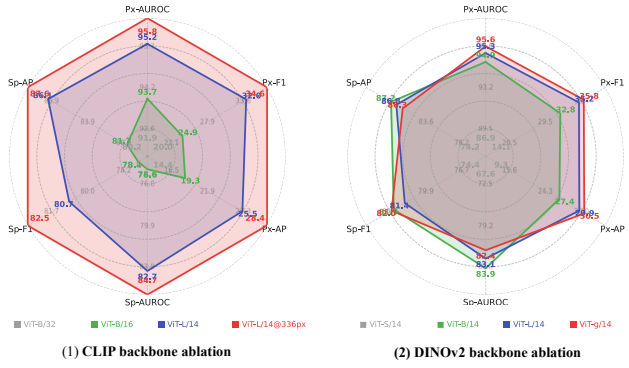
4.3. Ablation Studies

In this section, we use CLIP's ViT-L/14@336px backbone by default and conduct ablation studies on the VisA dataset to examine how different settings affect the proposed VisualAD method. We additionally include DINOv2-based ablations for part of the experiments, with the full results deferred to the appendix.

Table 2. Ablation study on different modules and loss components using ViT-L/14@336px. Both Image-level and Pixel-level metrics are reported as (AUROC, AP).

Ablation	SCA	SAF	Focal	Dice	Ctr	Image-level	Pixel-level
Module	✗	✓	✓	✓	✓	(82.3, 85.4)	(95.3, 27.4)
	✓	✗	✓	✓	✓	(50.5, 58.9)	(87.9, 3.5)
	✗	✗	✓	✓	✓	(48.0, 56.1)	(38.3, 0.8)
Loss	✓	✓	✗	✓	✓	(83.5, 86.8)	(95.3, 26.6)
	✓	✓	✓	✗	✓	(83.4, 86.5)	(95.3, 27.3)
	✓	✓	✓	✓	✗	(84.2, 87.0)	(95.4, 27.6)
VisualAD	✓	✓	✓	✓	✓	**(84.7, 87.6)**	**(95.8, 28.4)**

Effect of Different Components. We conduct separate ablations on modules and losses, with results summarized in Table 2. For the module ablation, removing any single component consistently degrades performance, indicating that each module contributes positively to the framework. Eliminating SCA weakens spatial alignment and fine-grained evidence aggregation, while removing the SAF undermines nonlinear recalibration of high-level semantics; dropping both leads to the most pronounced degradation, highlighting their complementary roles. For the loss ablation, excluding the Focal, Dice, or Ctr objective each yields a noticeable de-



(1) CLIP backbone ablation (2) DINOv2 backbone ablation

Figure 4. Radar plot comparing different pretrained ViT backbones under the ZSAD setting. We evaluate CLIP and DINO-based variants from small to large model scales. Larger backbones and higher input resolutions generally lead to consistent improvements in both pixel-level and sample-level metrics, with CLIP's ViT-L/14@336px giving the strongest sample-level accuracy and DINO's ViT-g/14 delivering the highest pixel-level precision.

Table 3. Ablation on the number of anchor queries using a ViT-L/14@336px backbone pre-trained with CLIP. Pixel-level and image-level metrics are reported as (AUROC, F_1-max, AP). The best and second-best results are highlighted in **bold** and underline, respectively.

Anchors	Pixel-level	Image-level
1	(95.2, 33.3, 27.0)	(84.7, 82.2, 87.3)
2	(95.2, 33.4, 27.2)	(83.8, 81.3, 86.7)
4	(**95.8**, **34.6**, **28.4**)	(84.7, 82.5, 87.6)
8	(95.0, 32.9, 26.8)	(84.0, 81.7, 86.7)
16	(95.1, 33.5, 26.9)	(**84.9**, **83.0**, **87.8**)
32	(94.6, 32.4, 26.1)	(84.5, 82.1, 87.6)

Table 4. Ablation on different layer combinations using ViT-L/14@336px backbone pre-trained with CLIP. Pixel-level and image-level metrics follow (AUROC, F_1-max, AP). Best and second-best results are in **bold** and underline.

Layers	Pixel-level	Image-level
{6}	(88.0, 11.9, 8.3)	(67.1, 75.0, 72.7)
{12}	(94.7, 28.4, 22.9)	(76.6, 77.3, 81.5)
{18}	(95.2, 31.3, 24.6)	(82.4, 80.5, 85.4)
{24}	(93.5, 25.9, 19.3)	(79.6, 79.1, 84.0)
{6, 12}	(93.7, 29.3, 24.1)	(79.5, 78.5, 83.6)
{6, 12, 18}	(95.0, 32.9, 27.1)	(83.4, 80.5, 86.4)
{6, 12, 18, 24}	(**95.8**, **34.6**, **28.4**)	(**84.7**, **82.5**, **87.6**)

cline, suggesting that the hybrid supervision balances classification focus, region consistency, and feature separability. Overall, the full model achieves the best image-level and pixel-level results, validating the effectiveness and synergy of all components.

Ablation on different pretrained ViT. We evaluate both CLIP and DINO pretrained ViTs from small to large capacities and observe a consistent trend: scaling the backbone and input resolution steadily improves pixel- and image-level performance, with gains tapering at the largest scales. Within CLIP, moving from ViT-B to ViT-L with 336px inputs yields the strongest overall results, while within DINO the larger variants approach or surpass smaller CLIP models in pixel-level F_1-max and AP. Notably, within the CLIP family ViT-L/14@336px attains the best sample-level results, whereas within the DINO family ViT-g/14 achieves the best pixel-level precision (highest pixel F_1-max). The Fig. 4 mirrors these patterns with broadly rising curves across AUROC, F_1-max, and AP.

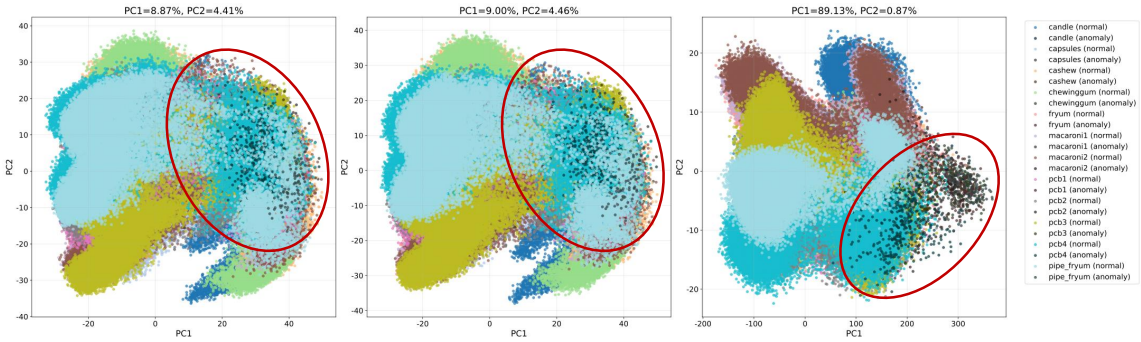
Influence of the number of anchors. Table 3 analyzes how the number of anchor queries in SCA affects model performance. At the pixel level, increasing the anchors from 1 to 4 yields steady improvements across AUROC, F_1-max, and AP, with $m = 4$ delivering the strongest overall results. Expanding the anchor set beyond 4 (e.g., 8–32) introduces redundancy and slightly degrades pixel-level accuracy, indicating that an excessively large anchor set dilutes spatial focus rather than enhancing it. Image-level metrics exhibit a flatter trend: while $m = 16$ provides the highest AUROC, F_1-max, and AP, the improvements over $m = 4$ remain within one to two percentage points. These findings suggest that a small anchor set is sufficient, and we adopt $m = 4$ as the default choice, achieving a balanced trade-off between pixel-level precision and image-level robustness.

Influence of ensemble of different patch-level image layers. Table 4 examines the effect of aggregating features from different intermediate layers of the frozen image encoder. Using a single layer ($\{6\}$, $\{12\}$, $\{18\}$, or $\{24\}$) shows that mid-level layers yield stronger representations than shallow or deep ones, with layer $\{18\}$ achieving the best balance between local detail and global semantics. When combining multiple layers, the performance improves consistently: $\{6,12,18\}$ already brings clear gains in both pixel- and sample-level metrics, and extending to $\{6,12,18,24\}$ further enhances stability and overall accuracy. These results demonstrate that multi-layer feature ensembles effectively capture complementary information across scales, leading to more robust anomaly localization and classification.

Table 5. Effect of positional modeling in the SCA module on VisA. Metrics are reported as (AUROC, F_1-max, AP). Best and second-best results are highlighted in **bold** and underline.

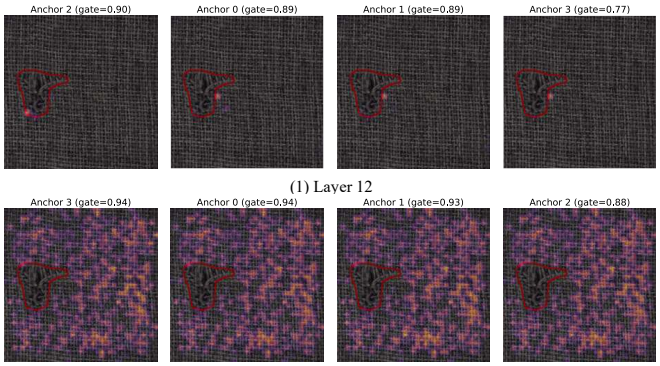
Positional modeling	Pixel-level	Image-level
None (no PE)	(93.0, 29.0, 22.5)	(82.4, 78.1, 84.5)
Learnable 1D Absolute	(**95.8**, **34.6**, **28.4**)	(**84.7**, **82.5**, **87.6**)
Fixed 1D Sinusoidal	(95.2, 33.8, 27.6)	(84.1, 81.9, 86.8)
Learnable 2D Absolute	(95.0, 33.2, 27.0)	(83.8, 81.6, 86.0)
Fixed 2D Sinusoidal	(94.6, 32.4, 26.1)	(83.4, 80.8, 85.4)
Learnable 2D Relative Bias	(94.4, 31.9, 25.7)	(83.0, 80.2, 84.9)

Influence of spatial positional modeling. Although we re-train the positional encoding associated with the inserted global tokens, we still observe that their interaction with patch tokens tends to blur the underlying spatial geometry and degrade pixel-level localization. To improve spatial consistency, we equip the SCA module with explicit positional modeling and conduct an ablation study comparing six variants: no positional encoding, learnable 1D absolute embeddings, fixed 1D sinusoidal encodings, learnable 2D absolute embeddings, fixed 2D sinusoidal encodings, and learnable 2D relative positional bias. As summarized in



(1) Original ViT (2) Add Normal and Abnormal Token (3) After SAF

Figure 5. PCA visualization of feature distributions under three configurations. From left to right: vanilla CLIP features, CLIP with normal/anomaly tokens, and CLIP with both tokens and the MLP-based transformation. The anomaly cluster (dark) becomes progressively more compact and moves farther from the normal cluster, while the variance concentrates along a single dominant axis, indicating increasingly discriminative representations for anomaly separation.



(1) Layer 12

(2) Layer 24

Figure 6. Layer-wise anchor attention maps on the *carpet* category. Anchors within each layer attend to similar regions, while different layers capture complementary cues: mid-level layers highlight fine-grained defect boundaries, and deeper layers emphasize broader normal regions, supporting accurate anomaly localization.

Table 5, we report both pixel-level and image-level performance in the format (AUROC, F_1-max, AP). A simple learnable 1D absolute embedding achieves the best overall trade-off across these metrics, while more complex 2D or relative formulations offer only marginal gains or even harm stability. Therefore, we adopt the learnable 1D absolute variant as the default configuration in the SCA module.

4.4. Visualization

We provide visual analyses of how anchor attention and feature representations evolve in VisualAD. Fig. 6 shows that anchors attend to highly consistent regions within each layer, indicating stable local evidence. Across layers, attention shifts from fine-grained defect boundaries in mid-level layers (e.g., layer 12) to broader, mostly normal regions in deeper layers (e.g., layer 24). This reflects the hierarchical behavior of the frozen ViT encoder and shows that the anchor-based SCA module effectively exploits complementary multi-layer cues for anomaly localization.

Fig. 5 shows the PCA distributions under three configurations. Vanilla CLIP features exhibit strong overlap between normal and anomalous samples, with PC1 explaining only 8.8% variance. Adding learnable normal and anomaly tokens slightly improves PC1 to 9.0%, suggesting a weak discriminative direction. After applying SAF, PC1 increases to 89.1%, yielding much clearer class separation and a more compact anomaly cluster. These results show that the tokens introduce initial discrimination, while SAF further enhances intra-class compactness and inter-class separation.

5. Conclusion

VisualAD introduces a purely visual approach to zero-shot anomaly detection by inserting two learnable global tokens (normal and abnormal) into a frozen Vision Transformer and grounding them with spatial evidence via Spatial Cross-Attention and a Self-Alignment Function. Trained only on auxiliary industrial data and without any text branch or cross-modal alignment, the method generalizes directly to unseen industrial and medical categories, delivering state-of-the-art image- and pixel-level performance. Ablations confirm that token learning, the SCA/SAF design, and multi-layer fusion each contribute indispensably to detection and localization. Future directions include adaptive layer selection, richer structured tokens for finer granularity, and lightweight uncertainty estimation to improve robustness under strong domain shift.

6. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62576001, No. 62206003).

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. 5, 6

[2] Ujjwal Baid, Satyam Ghodasara, and Michel Bilello. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *ArXiv*, abs/2107.02314, 2021. 5, 6

[3] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data*, 4, 2017. 5

[4] Babak Ehteshami Bejnordi, Mitko Veta, and Paul Johannes van Diest. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318:2199–2210, 2017. 5, 6

[5] Yoshua Bengio, L. Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. *ArXiv*, abs/1305.6663, 2013. 3

[6] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5, 6

[7] Jorge Bernal, Francisco Javier Sánchez, Glòria Fernández-Esparrach, Debora Gil, Cristina Rodríguez de Miguel, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 43:99–111, 2015. 5, 6

[8] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. *CoRR*, abs/2305.10724, 2023. 3

[9] 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, 2024. 2, 3, 5, 6

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 2, 3

[11] Xuhai Chen and et al. 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. *CoRR*, abs/2305.17382, 2023. 2, 5, 6

[12] Xuhai Chen, Jiangning Zhang, Guanzhong Tian, Haoyang He, Wuhao Zhang, Yabiao Wang, Chengjie Wang, Yunsheng Wu, and Yong Liu. CLIP-AD: A language-guided staged dual-path model for zero-shot anomaly detection. *CoRR*, abs/2311.00453, 2023. 2, 3, 5, 6

[13] Thomas Defard, Aleksandr Setkov, Angélique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *ICPR Workshops*, 2020. 3

[14] Alexey Dosovitskiy, Lucas Beyer, and Alexander Kolesnikov. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR*, 2021. 3

[15] Bin-Bin Gao. Metauas: Universal anomaly segmentation with one-prompt meta-learning. *ArXiv*, abs/2505.09265, 2025. 2

[16] Bin-Bin Gao. Learning to detect multi-class anomalies with just one normal image prompt. In *European Conference on Computer Vision*, 2025. 3

[17] 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*, abs/2505.09926, 2025. 2, 3

[18] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1705–1714, 2019. 3

[19] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pages 1705–1714. IEEE, 2019. 2

[20] Steven Hicks, Debesh Jha, Vajira Lasantha Thambawita, P. Halvorsen, Hugo Lewi Hammer, and M. Riegler. The endotect 2020 challenge: Evaluation and comparison of classification, segmentation and inference time for endoscopy. In *ICPR Workshops*, 2020. 5, 6

[21] Yanning Hou, Ke Xu, Junfa Li, Yanran Ruan, and Jianfeng Qiu. Enhancing zero-shot anomaly detection: Clip-sam collaboration with cascaded prompts. In *Chinese Conference on Pattern Recognition and Computer Vision*, 2024. 3

[22] Yanning Hou, Yanran Ruan, Junfa Li, Shanshan Wang, Jianfeng Qiu, and Ke Xu. Stackclip: Clustering-driven stacked prompt in zero-shot industrial anomaly detection. *ArXiv*, abs/2506.23577, 2025. 2

[23] Yanning Hou, Sihang Zhou, Ke Liang, Lingyuan Meng, Xiaoshu Chen, Ke Xu, Siwei Wang, Xinwang Liu, and Jian Huang. Soft reasoning paths for knowledge graph completion. In *International Joint Conference on Artificial Intelligence*, 2025. 3

[24] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 19606–19616. IEEE, 2023. 2, 3, 5, 6

[25] Debesh Jha, Pia Helen Smedsrud, M. Riegler, P. Halvorsen, Thomas de Lange, Dag Johansen, and Haavard D. Johansen. Kvasir-seg: A segmented polyp dataset. In *Conference on Multimedia Modeling*, 2019. 5, 6

[26] Daniel S. Kermany, Michael H. Goldbaum, and *et al. Identifying medical diagnoses and treatable diseases by image-based deep learning.* Cell*, 172:1122–1131.e9, 2018. 5, 6*

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. arXiv:2304.02643*, 2023. 3*

[28] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9659–9669, 2021. 3*

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 12888–12900, 2022. 2, 3*

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 19730–19742, 2023. 2*

[31] 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. 3*

[32] 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 (CVPR)*, pages 16838–16848, 2024. 2*

[33] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. CoRR*, 2023. 3*

[34] 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. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4744–4754, 2025. 3*

[35] Fabio Valerio Massoli, Fabrizio Falchi, Alperen Kantarci, Seymanur Akti, Hazim Kemal Ekenel, and Giuseppe Amato. MOCCA: multilayer one-class classification for anomaly detection. IEEE Trans. Neural Networks Learn. Syst.*, 33(6): 2313–2323, 2022. 2*

[36] Bjoern H Menze, András Jakab, Stefan Bauer, and et al. The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging*, 34:1993–2024, 2015. 5*

[37] 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 30th IEEE International Symposium on Industrial Electronics, ISIE 2021, Kyoto, Japan, June 20-23, 2021*, pages 1–6, 2021. 5, 6*

[38] 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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023. 1*

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. CoRR*, 2023. 2, 6*

[40] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2893–2901, 2019. 3*

[41] Zhen Qu, Xian Tao, Mukesh Prasad, Fei Shen, Zhengtao Zhang, Xinyi Gong, and Guiguang Ding. Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. In European Conference on Computer Vision*, 2024. 3*

[42] 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. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30398–30408, 2025. 3*

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, ICML*, pages 8748–8763, 2021. 2, 3, 6*

[44] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Scholkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14298–14308, 2021. 3*

[45] *Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter V. Gehler. Towards total recall in industrial anomaly detection. In* IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2022. 1*

[46] *Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection.* 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1829–1838, 2021. 3*

[47] *Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad Hossein Rohban, and Hamid R. Rabiee. Multiresolution knowledge distillation for anomaly detection.* 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14897–14907, 2020. 5, 6*

[48] *Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister. Learning and evaluating representations for deep one-class classification. In* International Conference on Learning Representations, ICLR*, 2021. 2*

[49] *Domen Tabernik, Samo Sela, Jure Skvarc, and Danijel Skocaj. Segmentation-based deep-learning approach for surface-defect detection.* J. Intell. Manuf.*, 31, 2020. 5, 6*

[50] *Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In* IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023. 1*

[51] *Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In* DAGM symposium in*, page 11, 2007. 5, 6*

[52] *Guoyang Xie, Jinbao Wang, Jiaqi Liu, Yaochu Jin, and Feng Zheng. Pushing the limits of fewshot anomaly detection in industry vision: Graphcore. In* The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023. 1*

[53] *Xudong Yan, Huaidong Zhang, Xuemiao Xu, Xiaowei Hu, and Pheng-Ann Heng. Learning semantic context from normal samples for unsupervised anomaly detection. In* AAAI Conference on Artificial Intelligence*, 2021. 3*

[54] *Haiming Yao, Yunkang Cao, Wei Luo, Weihang Zhang, Wenyong Yu, and Weiming Shen. Prior normality prompt transformer for multiclass industrial image anomaly detection.* IEEE Transactions on Industrial Informatics*, 20: 11866–11876, 2024. 3*

[55] *Xincheng Yao, Ruoqing Li, Zefeng Qian, Yan Luo, and Chongyang Zhang. Focus the discrepancy: Intra- and inter-correlation learning for image anomaly detection.* 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6780–6790, 2023. 3*

[56] *Jihun Yi and Sungroh Yoon. Patch SVDD: patch-level SVDD for anomaly detection and segmentation. In* 15th Asian Conference on Computer Vision ACCV*, pages 375–390, 2020. 2*

[57] *M. Zaheer, Jin ha Lee, M. Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm.* 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14171–14181, 2020. 3*

[58] *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 (ICLR)*, pages 1–33, 2024. 2, 3, 5, 6*

[59] *Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts.* 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17826–17836, 2024. 2*

[60] *Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In* Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv*, pages 392–408, 2022. 5, 6*