CLIP-FSAC: Boosting CLIP for Few-Shot Anomaly Classification with Synthetic Anomalies

Zuo Zuo[1,2] , **Yao Wu** [3] , **Baoqiang Li**[2] , **Jiahao Dong**[2] , **You Zhou**[4] , **Lei Zhou**[2] , **Yanyun Qu**[3*] and **Zongze Wu**[1,2,4*]

[1] National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[3]Xiamen University
[4]Shenzhen University
{Nostalgiaz}@stu.xjtu.edu.cn, {libaoqiang2023,dongjiahao2023}@email.szu.edu.cn,
yyqu@xmu.edu.cn, wuyao@stu.xmu.edu.cn, zzwu@szu.edu.cn

Abstract

Few-shot anomaly classification (FSAC) is a vital task in manufacturing industry. Recent methods focus on utilizing CLIP in zero/few normal shot anomaly detection instead of custom models. However, there is a lack of specific text prompts in anomaly classification and most of them ignore the modality gap between image and text. Meanwhile, there is distribution discrepancy between the pre-trained and the target data. To provide a remedy, in this paper, we propose a method to boost CLIP for few-normal-shot anomaly classification, dubbed CLIP-FSAC, which contains two-stage of training and alternating fine-tuning with two modality-specific adapters. Specifically, in the first stage, we train image adapter with text representation output from text encoder and introduce an image-to-text tuning to enhance multi-modal interaction and facilitate a better language-compatible visual representation. In the second stage, we freeze the image adapter to train the text adapter. Both of them are constrained by fusion-text contrastive loss. Comprehensive experiment results are provided for evaluating our method in few-normal-shot anomaly classification, which outperforms the state-of-the-art method by 12.2%, 10.9%, 10.4% AUROC on VisA for 1, 2, and 4-shot settings.

1 Introduction

Image anomaly detection (IAD) aims to detect anomalies in given images including classification (AC) and segmentation (AS) [Wu *et al.*, 2023]. As technology develops by leaps and bounds, anomaly detection based on deep learning is widely used in many fields such as intrusion detection, fraud detection and industrial inspection. Especially in industrial inspection, anomaly detection plays an important role which determines the quality of products. However, anomaly detection is a non-trivial task as abnormal images are scarce and anomalies are agnostic and diverse. Meanwhile, annotating anomaly regions is costly and labor-exhausting. To address these issues, more and more anomaly detection methods find anomalies in an unsupervised manner. In this manner, only anomaly-free samples are needed and there are no extra annotations. Training on massive normal samples, these methods can achieve an excellent performance.

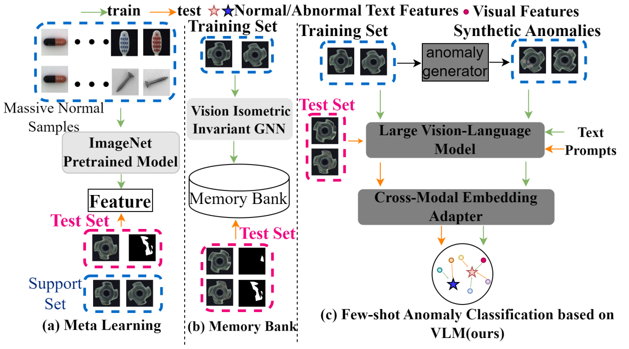


Figure 1: Different diagrams for few-shot anomaly detection. (a) few-shot unsupervised AC in meta learning. (b) few-shot unsupervised AC based on vision isometric invariant GNN using memory bank. Our proposed (c) leverages alignment capability between image and text of large vision-language model and fine-tune it for few-shot AD without extra memory bank and massive normal samples.

Nevertheless, data collecting in industrial scenarios is also a tricky challenge. With less training samples, the performance of these unsupervised methods will decrease substantially. Recently, some few-shot anomaly detection (FSAD) methods have been proposed. One prevailing paradigm is to solve FSAD by employing meta-learning [Huang *et al.*, 2022] (See Figure 1(a)). Since the difficulty in collecting massive similar normal images, Graphcore [Xie *et al.*, 2023] considers constructing a memory bank, which only needs a few normal images of target category (See Figure 1(b)). However, it neglects the burden brought by the additional inference time.

[*]Co-Correpsonding Authors.

Pre-trained vision-language models (VLMs) [Alayrac *et al.*, 2022; Radford *et al.*, 2021] have received ever-higher attention recently. In FSAC, WinCLIP [Jeong *et al.*, 2023] benefits from semantic connection and generalization ability of CLIP. It proposes a compositional prompt ensemble (CPE) on state words and prompt templates and aggregation of window/patch/image-level features aligned with texts. Inspired by WinCLIP, utilizing weight-pretrained visual encoder of CLIP as our training paradigm not only enables us to capture broader concepts with open vocabulary but also avoids introducing additional computational overheads brought by modules. However, directly matching representation from the original CLIP is sub-optimal in FSAD.

Text as domain-invariant prompt allows us to learn disentangled domain and category representation. It is crucial to design an optimal text prompt to capture visual-language relationship [Zhou *et al.*, 2022; Yao *et al.*, 2022]. In our earlier exploration, we change different combination of texts in zero-shot anomaly classification with CLIP and find that results vary widely. It demonstrates that it is difficult to find accurate prompts in anomaly classification. Meanwhile, there exists a large distribution discrepancy between industrial and natural images. Manual prompts cannot precisely disentangle the representations of domain and category. To this end, we propose a novel few-normal-shot anomaly classification method, dubbed CLIP-FSAC, focusing on leveraging the complementary advantage of cross-modality fusion to solve FSAC task.

Specifically, we adapt CLIP to identify anomalies via contrastive representation learning [Chen *et al.*, 2020]. Because of fewer normal samples, it is difficult to fine-tune image and text encoders of CLIP. Unlike learnable prompt in CoOp [Zhou *et al.*, 2022], we use two adapters to produce the target-oriented representations from image and text encoders of CLIP. To better transfer CLIP for FSAC, we propose a two-stage training strategy to train image and text adapters with normal and synthetic abnormal samples. Firstly we only optimize image adapter by our proposed fusion-text contrastive loss between adapted language-compatible visual features and text features from text encoder. After training image adapter, we add text adapter following text encoder and optimize it by similar contrastive restraint with frozen image adapter. Image-to-text cross-attention module between visual and text features is introduced to obtain visual-driven text features. Combining visual-driven text features with visual features can facilitate a better language-compatible visual representation. Then we exploit interaction between fused features and text features instead of using pure image features. Our main contributions are as follows:

- We empirically analyze that the designing of text prompts is important in zero-shot anomaly classification and propose CLIP-FSAC, which fully explores the potential of CLIP in few-shot anomaly classification.
- We propose a two-stage training strategy to optimize image adapter and text adapter respectively, which better discovers image-text matching capability of CLIP in few-shot anomaly classification.
- We propose to utilize visual-driven text features to enhance visual representations. Meanwhile, we use fusion-text matching task to replace image-text matching in original CLIP.
- Experimental results show superior performance of CLIP-FSAC in few-shot anomaly classification. CLIP-FSAC outperforms previous few-shot methods on VisA and MVTEC-AD for 1-shot, 2-shot and 4-shot, even surpassing many full-shot anomaly detection methods.

2 Related Work

2.1 Anomaly Detection

Anomaly detection[Akcay *et al.*, 2018] methods focus on unsupervised learning only using normal samples, which can be divided into three categories: reconstruction-based method, embedding-based method, and synthesizing-based method. Reconstruction-based methods [He *et al.*, 2023; Ristea *et al.*, 2022; Guo *et al.*, 2023] employ generative adversarial network (GAN) or diffusion models to identify anomalies through pixel-level comparisons between input images and their reconstructions. Embedding-based methods [Yao *et al.*, 2023; Tien *et al.*, 2023; Bae *et al.*, 2023] find anomalies by embedding features of normal samples into latent space and measuring feature-level deviations. While synthesizing-based methods [Zavrtanik *et al.*, 2021; Liu *et al.*, 2023b; Zhang *et al.*, 2023b] synthesize anomalies on normal images and train network with normal and abnormal samples. To address the scarcity of training samples, several few-shot anomaly detection methods have been proposed recently [Fang *et al.*, 2023]. RegAD uses meta-learning which is a new paradigm to train a category-agnostic anomaly detection model with mounts of samples of multiple other categories. GraphCore uses GNN to extract anomaly-free visual isometric invariant features to construct a memory bank with a few normal images. While we only need few normal samples of target category without memory bank.

2.2 Vision-Language Models

Vision-Language Models (VLMs) [Zhu *et al.*, 2023; Jia *et al.*, 2021] which explore the interaction between textual and visual representation witness significant advancements. VLMs have been successfully transferred in many downstream tasks [Wang *et al.*, 2023; Li *et al.*, 2023], especially CLIP [Radford *et al.*, 2021] trained on a dataset consisted of 400 million image-text pairs which is commonly used pre-trained vision-language models. TCM [Yu *et al.*, 2023] turns CLIP directly for text detection without pre-training. CLIP-LIT [Liang *et al.*, 2023] explores the potential of CLIP for pixel-level image enhancement. DenseCLIP [Rao *et al.*, 2022] applies CLIP in dense prediction task. Recently, some works propose to use CLIP for anomaly detection. WinCLIP [Jeong *et al.*, 2023] uses CLIP for zero-/few-Shot anomaly classification and segmentation without fine-tuning, which designs a set of prompts and proposes Window-based CLIP which divides image into patches to extract dense visual features. Then it detects anomalies via vision-language alignment. But WinCLIP is too dependent on original representation of CLIP, as its architecture is restricted to CLIP. Therefore we aim to exploit the semantic relationships between task-specific images and its associated texts.

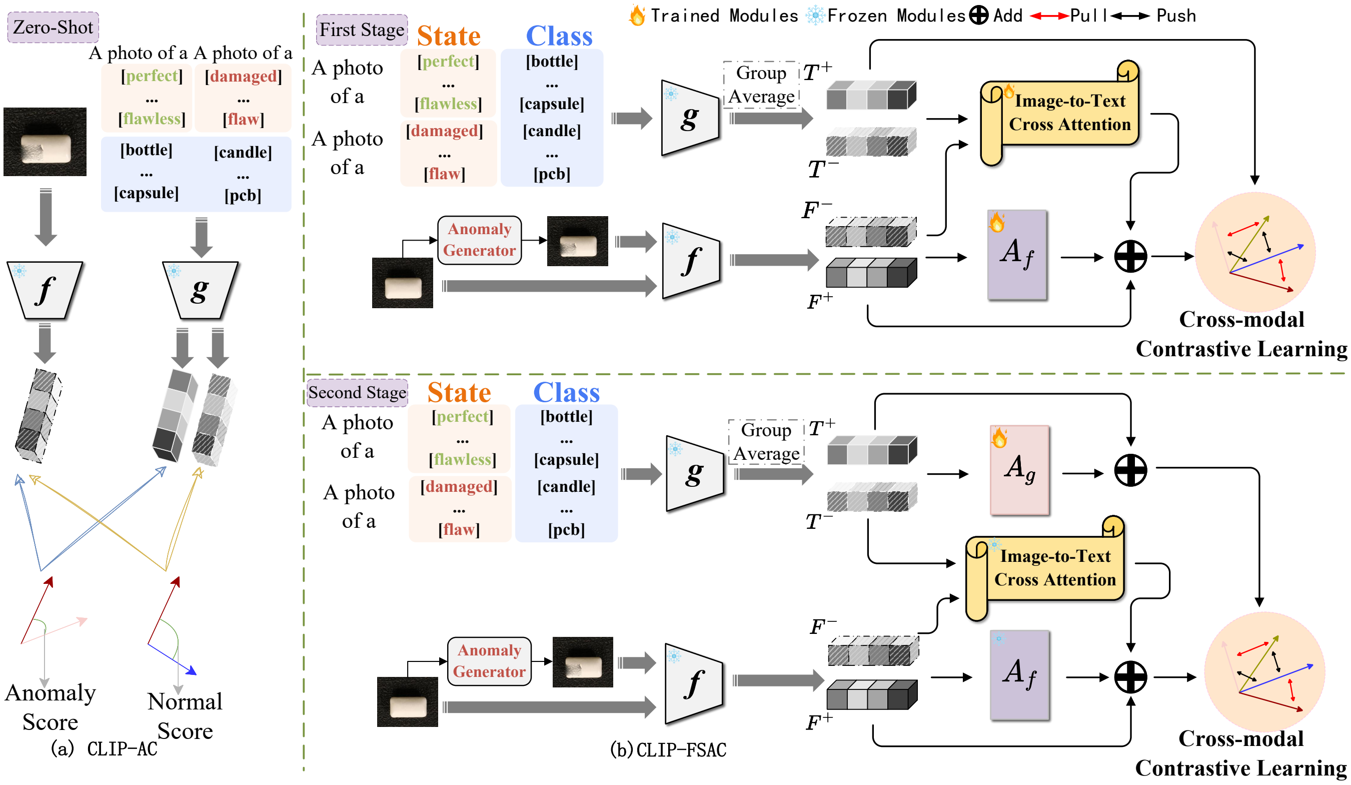


Figure 2: The framework of CLIP-FSAC. Zero-shot indicates zero-shot anomaly classification with original CLIP. f and g are image and text encoders of CLIP, A_f and A_g are image and text adapter.

3 Method

3.1 Preliminaries: Zero-Shot Anomaly Classification with CLIP

Contrastive language image pre-training (CLIP) can effectively learn joint image-language representations which consists of two encoders, an image encoder $f(\cdot)$ and a text encoder $g(\cdot)$. Given an image, CLIP can find its corresponding language representation. This ability of image-text matching can be directly used in anomaly classification. WinCLIP proposes a zero-shot anomaly classification framework named CLIP-AC illustrated in Figure 2(a). In CLIP-AC, two class prompts, s^+ =“normal [o]” and s^- =“anomalous [o]” are used to identify anomaly images. [o] is a simple object-level label. For a test image, its anomaly score is defined as consine similarity of image and abnormal text embeddings. To further improve performance, WinCLIP also proposes compositional prompt ensemble (CPE) including combinations of pre-defined lists of state words per label and text templates.

3.2 Overview of CLIP-FSAC

We propose CLIP-FSAC, which further pushes the upper limits of CLIP in anomaly classification task. The overall pipeline is shown in Figure 2(b). Our scheme is built upon CLIP and compositional prompt ensemble with two adapters for distributing representations from original CLIP via two-stage training strategy. Before fine-tuning, we generate synthetic anomaly samples based on given normal images. Then we feed positive and negative samples into frozen image and text encoders to obtain visual and text features. In the first training stage, we only use image adapter to transfer visual embeddings and use designed image-to-text cross-attention module to obtain visual-driven text features which are added to transferred visual features in the following. Image adapter and cross-attention module are optimized by fusion-text contrastive loss between cross-modality fused features and text features. In the second training stage, we freeze image adapter and cross-attention module and optimize text adapter using the same loss as in the first stage. Significantly, we only need a few normal images less than 4 for fine-tuning.

3.3 Anomalies Generation

To optimize image and text adapters via contrastive learning, only a few anomaly-free images are not enough as lack of negative samples. Therefore, given k-shot normal samples, we generate synthetic anomaly images as negative samples like data-augmentation in few-shot learning. Due to various distributions of anomalies, we mainly use two methods for generating anomalies to simulate natural sub-image irregularities in two different industrial datasets. The first way to synthesize anomalies is via random perturbation [Cao *et al.*, 2023]. Specifically, we randomly choose some square regions in normal images and replace these regions with random values sampled from a Gaussian normal distribution as shown in Figure 3(a). The second approach proposed by NSA [Schlüter *et al.*, 2022] is more complicated than the first

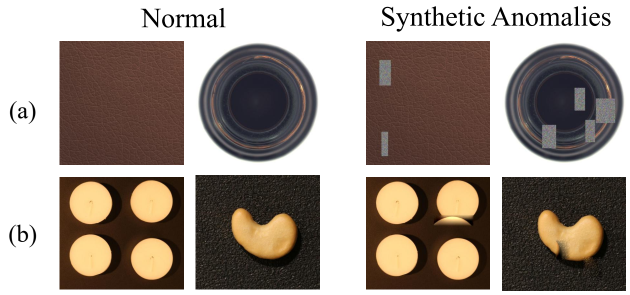


Figure 3: Synthetic anomalies. (a) random perturbation. (b) NSA.

method. NSA integrates Poisson image editing and Gamma-distribution-based patch shape sampling strategy in anomalies generation process to generate more natural and diverse anomalies as illustrated in Figure 3(b) which are necessary for better performance in testing phase. We denote normal image as x_i^+ and synthetic anomaly image as $x_{i,j}^-$, where (i,j) represents the j anomaly image based on x_i^+. For each normal image, we will generate more than one anomaly images

$$x_{i,j}^- = Anomaly\,Generator(x_i^+). \tag{1}$$

3.4 CLIP-FSAC

Adapters

Given a normal and abnormal image pair x_i^+ and $x_{i,j}^-$, the [CLS] token embeddings of image features are denoted as

$$F_i^+ = f(x_i^+), \quad F_{i,j}^- = f(x_{i,j}^-), F_i^+, F_{i,j}^- \in \mathbb{R}^{1\times C}. \tag{2}$$

With normal and abnormal compositional prompts, we compute the average of [EOS] tokens encoded by text encoder $g(\cdot)$ which are treated as text representations to represent the normal and anomalous samples, denoted as T^+ and T^-, $T^+, T^- \in \mathbb{R}^{1\times C}$. C is tokens dimension. To better exploit the image-text matching ability of CLIP in anomaly classification, we use two adapters denoted as $A_f(\cdot)$ and $A_g(\cdot)$ for image and text adaptation respectively. To avoid loss of prior information of CLIP, we blend original knowledge and adaptive knowledge via residual connection with residual ratio α and β controlling the proportion of the original and adapted representations [He *et al.*, 2016; Gao *et al.*, 2021], AF and AT denote image and text adaptive features which can be written as:

$$\begin{aligned} AF_i^+ &= \alpha_1 * A_f(F_i^+) + \alpha_2 * F_i^+,\ AF_i^+ \in \mathbb{R}^{1\times C} \\ AF_{i,j}^- &= \alpha_1 * A_f(F_{i,j}^-) + \alpha_2 * F_{i,j}^-, AF_{i,j}^- \in \mathbb{R}^{1\times C}, \end{aligned} \tag{3}$$

$$\begin{aligned} AT^+ &= \beta_1 * A_g(T^+) + \beta_2 * T^+, AT^+ \in \mathbb{R}^{1\times C} \\ AT^- &= \beta_1 * A_g(T^-) + \beta_2 * T^-, AT^- \in \mathbb{R}^{1\times C}. \end{aligned} \tag{4}$$

To enhance the visual representations and mine most correlated visual cues, we introduce image-to-text cross-attention module in image adapter to obtain visual-driven text features IT_i^+. We concatenate AT^+ and AT^- as compositional text features denoted as $\psi \in \mathbb{R}^{2\times C}$. Then we calculate cross attention between compositional text feature and positive visual embedding as

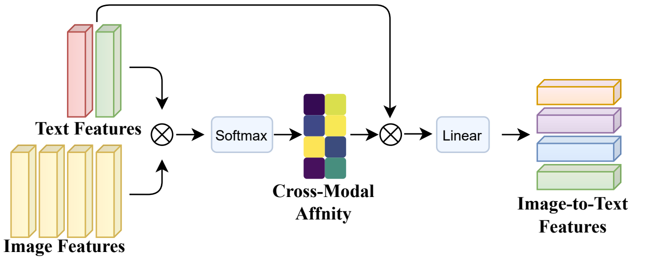


Figure 4: Architecture of image-to-text cross attention.

$$Atten^+ = Linear(Softmax(\frac{F_i^+ \cdot \psi^{\mathrm{T}}}{\sqrt{C}})\psi), \tag{5}$$

where $Linear(\cdot)$ is a fully-connected layer. $Atten^+ \in \mathbb{R}^{1\times C}$ is positive visual-driven text features. $Atten^- \in \mathbb{R}^{1\times C}$ is negative one obtained by the same way of $Atten^+$.

The structure of cross-attention module is illustrated in Figure 4. We add these two visual-driven text features to adapted visual embeddings AF_i^+ and $AF_{i,j}^-$ to generate cross-modality fused features instead of pure visual features for normal and abnormal classes, denoted as CF_i^+ and $CF_{i,j}^-$:

$$CF_i^+ = AF_i^+ + Atten^+, CF_{i,j}^- = AF_{i,j}^- + Atten^-. \tag{6}$$

We optimize adapters using contrastive representation learning between cross-modality fused features and text features.

Training

CLIP-FSAC contains two-stage of training to optimize image and text adapters respectively. We denote the training set and test set as χ_{train} including less than 4 normal samples and χ_{test}. In the first training stage, we only optimize image adapter $A_f(\cdot)$. Given a normal image $x_i^+ \in \chi_{train}$, we first generate an anomalous image $x_{i,j}^-$ as its counterpart sample. With designed text prompts and a pair of positive and negative samples, we train image adapter $A_f(\cdot)$ using fusion-text contrastive loss. After training image adapter, we only optimize parameters in text adapter $A_g(\cdot)$ using the same loss as in the first stage. Our proposed fusion-text contrastive loss consists of two parts. The first part is fusion-to-text contrastive loss $\mathcal{L}_{f2t}$ which is calculated as:

$$\begin{aligned} \mathcal{L}_{f2t} = &-log\frac{exp(s(CF_i^+, AT^+))}{exp(s(CF_i^+, AT^+)) + exp(s(CF_i^+, AT^-))} \\ &-log\frac{exp(s(CF_{i,j}^-, AT^-))}{exp(s(CF_{i,j}^-, AT^-)) + exp(s(CF_{i,j}^-, AT^+))}, \end{aligned} \tag{7}$$

Setup	Method	Venue	VisA			MVTEC-AD		
			I-AUROC	I-AUPR	F1-MAX	I-AUROC	I-AUPR	F1-MAX
1-shot	PaDiM [Defard *et al.*, 2021]	ICPRW2021	62.8	68.3	75.3	76.6	88.1	88.2
	PatchCore [Roth *et al.*, 2022]	CVPR2022	79.9	82.8	81.7	83.4	92.2	90.5
	RegAD [Huang *et al.*, 2022]	ECCV2022	-	-	-	82.4	-	-
	GraphCore [Xie *et al.*, 2023]	ICLR2023	-	-	-	89.9	-	-
	WinCLIP [Jeong *et al.*, 2023]	CVPR2023	83.8	85.1	83.1	93.1	96.5	93.7
	CLIP-FSAC (ours/one-stage)		81.1	84.1	80.0	93.2	96.7	94.1
	CLIP-FSAC (ours/two-stage)		96.0$_{\uparrow 12.2}$	96.0$_{\uparrow 10.9}$	93.4$_{\uparrow 10.3}$	95.5$_{\uparrow 2.4}$	97.5$_{\uparrow 1.0}$	95.0$_{\uparrow 1.3}$
2-shot	PaDiM [Defard *et al.*, 2021]	ICPRW2021	67.4	71.6	75.7	78.9	89.3	89.2
	PatchCore [Roth *et al.*, 2022]	CVPR2022	81.6	84.8	82.5	86.3	93.8	92.0
	RegAD [Huang *et al.*, 2022]	ECCV2022	-	-	-	85.7	-	-
	GraphCore [Xie *et al.*, 2023]	ICLR2023	-	-	-	91.9	-	-
	WinCLIP [Jeong *et al.*, 2023]	CVPR2023	84.6	85.8	83.0	94.4	97.0	94.4
	CLIP-FSAC (ours/one-stage)		80.4	83.3	79.5	93.1	96.7	94.0
	CLIP-FSAC (ours/two-stage)		95.5$_{\uparrow 10.9}$	95.5$_{\uparrow 9.7}$	93.3$_{\uparrow 10.3}$	92.6$_{\downarrow 1.8}$	95.4$_{\downarrow 1.6}$	93.0$_{\downarrow 1.4}$
4-shot	PaDiM [Defard *et al.*, 2021]	ICPRW2021	72.8	75.6	78.0	80.4	90.5	90.2
	PatchCore [Roth *et al.*, 2022]	CVPR2022	85.3	87.5	84.3	88.8	94.5	92.6
	RegAD [Huang *et al.*, 2022]	ECCV2022	-	-	-	88.2	-	-
	GraphCore [Xie *et al.*, 2023]	ICLR2023	-	-	-	92.9	-	-
	WinCLIP [Jeong *et al.*, 2023]	CVPR2023	87.3	88.8	84.2	95.2	97.3	94.7
	CLIP-FSAC (ours/one-stage)		81.3	84.1	80.1	93.0	96.7	94.0
	CLIP-FSAC (ours/two-stage)		97.7$_{\uparrow 10.4}$	97.9$_{\uparrow 9.1}$	95.5$_{\uparrow 11.3}$	95.8$_{\uparrow 0.6}$	98.0$_{\uparrow 0.7}$	95.6$_{\uparrow 0.9}$

Table 1: Anomaly Classification performance comparison of the proposed CLIP-FSAC against the SOTA approaches on the VisA and MVTEC-AD benchmarks. One stage refers to the results trained only on the image adapter, while two-stage refers to the results obtained by freezing the image adapter and training the text adapter. Red indicates the best result, and blue displays the second-best.

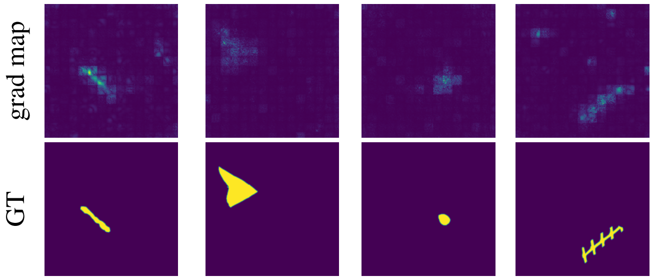


Figure 5: Visualization of grad maps and ground truth. Yellow regions in GT denote anomalies.

$s(\cdot,\cdot)$ is consine similarity function. The other part is text-to-fusion contrastive loss $\mathcal{L}_{t2f}$. It is calculated as:

$$\mathcal{L}_{t2f} = -log\frac{exp(s(CF_i^+, AT^+))}{exp(s(CF_i^+, AT^+)) + exp(s(CF_{i,j}^-, AT^+))} \\ -log\frac{exp(s(CF_{i,j}^-, AT^-))}{exp(s(CF_{i,j}^-, AT^-)) + exp(s(CF_i^+, AT^-))}. \tag{8}$$

Fusion-text contrastive loss $\mathcal{L}_{con}$ used in our method is a combination of fusion-to-text and text-to-fusion contrastive loss

$$\mathcal{L}_{con} = \frac{1}{2}(\mathcal{L}_{f2t} + \mathcal{L}_{t2f}). \tag{9}$$

3.5 Anomaly Classification with CLIP-FSAC

Anomalies Generator is no longer needed during testing and CLIP-FSAC is an end-to-end anomaly classification method. For a test image $x_{test} \in \chi_{test}$, we feed it into CLIP image encoder and adapter to obtain cross-modality fused features CF_{test}. With specific normal and abnormal text representations AT^+ and AT^-, we can calculate positive score S^+ and negative score S^- for given testing image x_{test} as follows:

$$S^+(x_{test}) = \frac{exp(\frac{s(CF_{test}, AT^+)}{\tau})}{exp(\frac{s(CF_{test}, AT^-)}{\tau}) + exp(\frac{s(CF_{test}, AT^+)}{\tau})}, \tag{10}$$

$$S^-(x_{test}) = \frac{exp(\frac{s(CF_{test}, AT^-)}{\tau})}{exp(\frac{s(CF_{test}, AT^-)}{\tau}) + exp(\frac{s(CF_{test}, AT^+)}{\tau})}, \tag{11}$$

where τ is a temperature hyper-parameter. Finally, we calculate the final anomaly score as

$$AS(x_{test}) = \frac{S^-}{S^- + S^+}. \tag{12}$$

Notably, we calculate the absolute value of gradients $grad = \left|\frac{\partial AS(x_{test})}{\partial x_{test}}\right| \in \mathbb{R}^{H \times W \times 3}$ via back propagation. Then we calculate the mean value of $grad$ along channel dimension as $grad\_map$. After visualizing $grad\_map$ as illustrated in Figure 5, we find that anomalies regions in image x_{test} contribute more to final anomaly score $AS(x_{test})$ than

Method	k-shot	VisA	MVTEC
CLIP-FSAC(ours)	1-shot	96.0	95.5
CLIP-FSAC(ours)	2-shot	95.5	92.6
CLIP-FSAC(ours)	4-shot	97.7	95.8
EfficientAD-S [Batzner *et al.*, 2024]	full-shot	97.5	98.8
FAIR [Liu *et al.*, 2023a]	full-shot	96.7	98.6
EdgRec [Liu *et al.*, 2022]	full-shot	94.2	97.8
CutPaste [Li *et al.*, 2021]	full-shot	-	95.2
PaDiM [Defard *et al.*, 2021]	full-shot	-	95.8
MKD [Salehi *et al.*, 2021]	full-shot	-	87.7

Table 2: Quantitative comparison (I-AUROC) between CLIP-FSAC and full-shot methods on VisA and MVTEC-AD datasets.

Training Strategy	Joint Training	Two-Stage Training
1-shot	94.5/94.9/90.5	96.0/96.0/93.4
2-shot	95.7/96.1/92.7	95.5/95.5/93.3
4-shot	96.7/97.4/93.3	97.7/97.9/95.5

Table 3: Ablation study on VisA of the effectiveness of two-stage training strategy. (I-AUROC/I-AUPR/F1-MAX)

Dataset	Setup	w/o CA	w/ CA
VisA	K=1	86.3/87.1/86.0	96.0/96.0/93.4
	K=2	86.0/87.3/86.8	95.5/95.5/93.3
	K=4	92.1/93.0/90.8	97.7/97.9/95.5
MVTEC-AD	K=1	89.9/94.0/93.2	95.5/97.5/95.0
	K=2	89.4/94.7/91.8	92.6/95.4/93.0
	K=4	95.1/97.4/94.7	95.8/98.0/95.6

Table 4: Ablation analysis of the proposed image-to-text cross attention module. (I-AUROC/I-AUPR/F1-MAX)

Dataset	Setup	Random perturbation	NSA
VisA	K=1	97.4/97.9/94.3	96.0/96.0/93.4
	K=2	98.2/98.6/95.5	95.5/95.5/93.3
	K=4	98.2/98.5/95.2	97.7/97.9/95.5
MVTEC-AD	K=1	95.5/97.5/95.0	90.5/95.0/92.6
	K=2	92.6/95.4/93.0	94.4/97.2/94.5
	K=4	95.8/98.0/95.6	95.6/98.3/95.5

Table 5: Ablation studies on different methods of synthesizing anomalies. (I-AUROC/I-AUPR/F1-MAX)

normal regions. This demonstrates cross-modal description ability of CLIP and explains why our method based on CLIP can give a good performance in anomaly classification.

4 Experiments

4.1 Experiment Setup

Dataset

Our experiments are conducted on both MVTEC-AD [Bergmann *et al.*, 2019] and VisA [Zou *et al.*, 2022] datasets. MVTEC-AD includes 15 classes including 5 different texture categories and 10 different object categories with totally 5354 high-resolution images. VisA covers 12 objects such as in 3 domain consisting of 10,821 high-resolution color images.

Evaluation Metrics

Our few-shot anomaly classification performance is evaluated based on three metrics following previous anomaly detection methods: (1) Area Under the Receiver Operating Characteristic (I-AUROC) which is the most commonly used metric for anomaly detection tasks, (2) Area Under the Precision-Recall curve (I-AUPR) which is proposed to address imbalance issues [Zou *et al.*, 2022] and (3) F1-Max which measures F1-score for anomaly classification at optimal threshold.

Implementation Details

We utilize OpenCLIP and load their pre-trained checkpoints called LAION-400M [Schuhmann *et al.*, 2021] into image and text encoders like WinCLIP. Meanwhile, we directly use CPE proposed in WinCLIP. In our method, we use random perturbation to generate anomalies on Mvtec AD and NSA on VisA. Our image adapter consists of 2-layer multi-layer perceptron (MLP) and text adapter is composed of one MLP. We use Adam optimizer with learning rate of 0.0005 for image adapter and 0.0001 for text adapter. Training epoch is set to 100 for all datasets. Our few-shot anomaly classification settings are set to 1-shot, 2-shot and 4-shot. So batch size is set to the number of training samples. α_1 and α_2 in Equation 3 are both set to 1 for MVTEC-AD and VisA. β_1 and β_2 in Equation 4 are set 0.1 and 0.9 respectively for MVTEC-AD dataset but both are set 1 for VisA.

4.2 Comparison to the State-of-the-Arts

The results of CLIP-FSAC on MVTEC-AD and VisA datasets are shown in Table 1. We compare our anomaly classification performance with previous state-of-the-art anomaly detection methods. In the few-normal-shot setup, our CLIP-FSAC achieves new SOTA of 96.0%, 95.5% and 97.7% of I-AUROC on VisA for 1-shot, 2-shot and 4-shot respectively, improving upon the state-of-the-art WinCLIP by 12.2%, 10.9% and 10.4% on 1-shot, 2-shot and 4-shot considerably. For MVTEC AD dataset, CLIP-FSAC reaches new SOTA of 95.5%, 95.8% of I-AUROC on MVTEC AD for 1-shot and 4-shot, surpassing WinCLIP by 2.4% and 0.6%. We note that CLIP-FSAC outperforms almost previous state-of-the-art anomaly detection methods. To measure the performance of our method comprehensively, we also use two other metrics I-AUPR and F1-Max to demonstrate the classification results except I-AUROC. As shown in Table 1, I-AUPR and F1-Max of CLIP-FSAC also surpass many previous SOTA methods on MVTEC AD and VisA datasets.

4.3 Comparison with Full-Shot Methods

As illustrated in Table 2, we compare our results in few-shot setup with previous full-shot methods on VisA and MVTEC

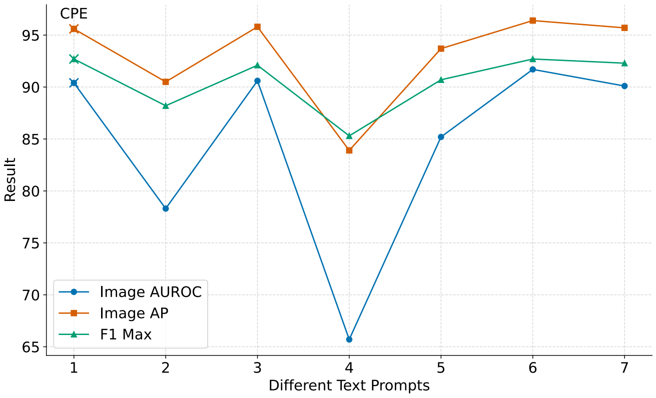


Figure 6: Performance comparison of different text prompts. CPE is compositional prompt ensemble proposed in WinCLIP.

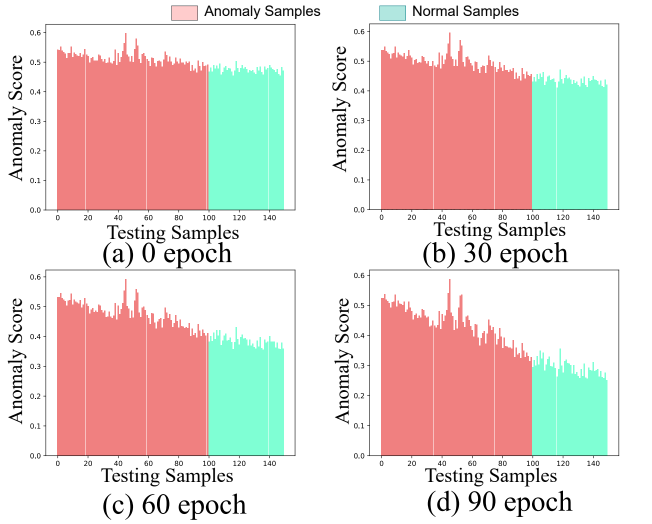
AD datasets. Though with less normal samples, our classification results are competitive with these full-shot methods. The state-of-the-art method on VisA is DiffusionAD [Zhang *et al.*, 2023a] which can reach 98.8% I-AUROC of using full-shot samples. But I-AUROC of CLIP-FSAC achieves 97.7% which is very close to 98.8% I-AUROC of DiffusionAD. It outperforms most full-shot methods only using four normal training samples on VisA. On MVTEC-AD dataset, we can reach I-AUROC score of 95.8% with 4-shot surpassing many full-shot and few-shot anomaly detection methods.

4.4 Ablation Studies

Effect of two-stage training strategy. Better aligning text and image embeddings is vital for FSAC. There are two approaches to train image and text adapters. The first way is to train image and text adapters simultaneously called joint training strategy. The other way we proposed is to train them separately. After training image adapter, its parameters are frozen and text adapter is trained. We compare the performance of these two training strategies. As illustrated in Table 3, the two-stage training is more effective, improve more than 1% I-AUROC in 1-shot and 4-shot setup on VisA.

Effect of cross attention module. Compare with image-text matching, we use image-to-text cross-attention module to generate visual-driven text features which are used to enhance image embeddings. Then we optimize adapters via aligning cross-modality and image embeddings. In Table 4, we evaluate the effectiveness of cross-attention module on MVTEC-AD and VisA. We find that the performance with cross attention module is better on both datasets in any setup. Notably, it improves almost 10% I-AUROC on VisA in 1-shot. This observation suggests that there is a better alignment between cross-modality fused features and text features.

Effect of synthetic anomalies. In ablation studies, we investigate the influence of methods in synthesizing anomalies. Because synthetic anomalies are used for training, we hope that they can simulate realistic anomalies as far as possible. We mainly use random perturbation and NSA to generate anomaly samples. So we use these two approaches for both MVTEC-AD and VisA. The experimental results are shown in Table 5, which indicate that the upper limit of random perturbation is higher than NSA, which surpasses 98% I-AUROC on VisA. But the performance of NSA is better than random perturbation on MVTEC-AD in 2-shot.



(a) 0 epoch
(b) 30 epoch
(c) 60 epoch
(d) 90 epoch

Figure 7: Anomaly score distribution of testing samples in different epochs: (a) 0 epoch, (b) 30 epoch, (c) 60 epoch and (d) 90 epoch. X-axis: testing samples, Y-axis:anomaly score.

Performance comparison of different text prompts. We try to change text prompts in zero-shot anomaly classification based on CLIP. Results vary a lot in different text prompts as shown in Figure 6. CPE means compositional prompt ensemble proposed in WinCLIP. We find that the performance of the sixth group of prompts even surpasses CPE. This demonstrates that it is difficult to design an optimal text prompts.

Visualization of anomaly score distribution. Finally, we visualize the anomaly score distribution of testing samples depicted by histograms. As can be seen in Figure 7, discrepancies between anomaly scores of normal samples and anomaly samples become more and more obvious as training epoch increases. It demonstrates that our CLIP-FSAC is effective and adapts CLIP for better performance in few-shot anomaly classification task.

5 Conclusion

In this paper, we propose a novel fine-tuning framework named CLIP-FSAC to adapt CLIP for few-shot anomaly classification. We use two adapters to embed image and text features into a space for better alignment. Meanwhile, we propose a image-to-text cross-attention module to enhance visual embeddings with visual-driven text features. The adapters and cross-attention module are optimized via two-stage training strategy. Our method achieves state-of-the-art results on VisA and MVTEC-AD datasets, even outperforming many full-shot methods. We believe that cross-modal description ability of the vision-language pre-training model still have a lot of room for improvement in zero-shot anomaly classification. In the future, we will further explore vision-language pre-training model and enhance its alignment capability between texts and images for anomaly classification tasks.

Acknowledgments

This work was supported in part by Guangdong Major Project of Basic and Applied Basic Research under Grant 2023B0303000009 and National Major Scientific Instruments and Equipments Development Project of National Natural Science Foundation of China under Grant 62327808.

References

[Akcay *et al.*, 2018] Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *ACCV*, pages 622–637, 2018.

[Alayrac *et al.*, 2022] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *NIPS*, pages 23716–23736, 2022.

[Bae *et al.*, 2023] Jaehyeok Bae, Jae-Han Lee, and Seyun Kim. Pni : Industrial anomaly detection using position and neighborhood information. In *ICCV*, pages 6373–6383, 2023.

[Batzner *et al.*, 2024] Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In *WACV*, pages 128–138, 2024.

[Bergmann *et al.*, 2019] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, pages 9584–9592, 2019.

[Cao *et al.*, 2023] Yunkang Cao, Xiaohao Xu, Zhaoge Liu, and Weiming Shen. Collaborative discrepancy optimization for reliable image anomaly localization. *TII*, 19(11):10674–10683, 2023.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, ICML'20, 2020.

[Defard *et al.*, 2021] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In *ICPRW*, page 475–489, 2021.

[Fang *et al.*, 2023] Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *ICCV*, pages 17481–17490, 2023.

[Gao *et al.*, 2021] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv*, 2021.

[Guo *et al.*, 2023] Hewei Guo, Liping Ren, Jingjing Fu, Yuwang Wang, Zhizheng Zhang, Cuiling Lan, Haoqian Wang, and Xinwen Hou. Template-guided hierarchical feature restoration for anomaly detection. In *ICCV*, pages 6447–6458, 2023.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[He *et al.*, 2023] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. Diad: A diffusion-based framework for multi-class anomaly detection. *arXiv*, 2023.

[Huang *et al.*, 2022] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *ECCV*, page 303–319, 2022.

[Jeong *et al.*, 2023] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, pages 19606–19616, 2023.

[Jia *et al.*, 2021] 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 *ICML*, pages 4904–4916, 2021.

[Li *et al.*, 2021] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *CVPR*, pages 9659–9669, 2021.

[Li *et al.*, 2023] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels. In *AAAI*, pages 1405–1413, 2023.

[Liang *et al.*, 2023] Zhexin Liang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Iterative prompt learning for unsupervised backlit image enhancement. In *ICCV*, pages 8094–8103, 2023.

[Liu *et al.*, 2022] Tongkun Liu, Bing Li, Zhuo Zhao, Xiao Du, Bingke Jiang, and Leqi Geng. Reconstruction from edge image combined with color and gradient difference for industrial surface anomaly detection. *arXiv*, 2022.

[Liu *et al.*, 2023a] Tongkun Liu, Bing Li, Xiao Du, Bingke Jiang, Leqi Geng, Feiyang Wang, and Zhuo Zhao. Fair: Frequency-aware image restoration for industrial visual anomaly detection. *arXiv*, 2023.

[Liu *et al.*, 2023b] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *CVPR*, pages 20402–20411, 2023.

[Radford *et al.*, 2021] 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 *ICML*, pages 8748–8763, 2021.

[Rao *et al.*, 2022] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, pages 18061–18070, 2022.

[Ristea *et al.*, 2022] Nicolae-Cătălin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B. Moeslund, and Mubarak Shah. Self-supervised predictive convolutional attentive block for anomaly detection. In *CVPR*, pages 13566–13576, 2022.

[Roth *et al.*, 2022] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, pages 14298–14308, 2022.

[Salehi *et al.*, 2021] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid R. Rabiee. Multiresolution knowledge distillation for anomaly detection. In *CVPR*, pages 14897–14907, 2021.

[Schlüter *et al.*, 2022] Hannah M. Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *ECCV*, page 474–489, 2022.

[Schuhmann *et al.*, 2021] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, 2021.

[Tien *et al.*, 2023] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan T.M. Duong, Chanh D. Tr. Nguyen, and Steven Q. H. Truong. Revisiting reverse distillation for anomaly detection. In *CVPR*, pages 24511–24520, 2023.

[Wang *et al.*, 2023] Henan Wang, Muli Yang, Kun Wei, and Cheng Deng. Hierarchical prompt learning for compositional zero-shot recognition. In *IJCAI*, pages 1470–1478, 2023.

[Wu *et al.*, 2023] Yao Wu, Mingwei Xing, Yachao Zhang, Yuan Xie, Zongze Wu, and Yanyun Qu. Perturbed progressive learning for semisupervised defect segmentation. *TNNLS*, 2023.

[Xie *et al.*, 2023] Guoyang Xie, Jinbao Wang, Jiaqi Liu, Yaochu Jin, and Feng Zheng. Pushing the limits of few-shot anomaly detection in industry vision: Graphcore. In *ICLR*, 2023.

[Yao *et al.*, 2022] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv*, 2022.

[Yao *et al.*, 2023] Xincheng Yao, Ruoqi Li, Zefeng Qian, Yan Luo, and Chongyang Zhang. Focus the discrepancy: Intra- and inter-correlation learning for image anomaly detection. In *ICCV*, pages 6803–6813, 2023.

[Yu *et al.*, 2023] Wenwen Yu, Yuliang Liu, Wei Hua, Deqiang Jiang, Bo Ren, and Xiang Bai. Turning a clip model into a scene text detector. In *CVPR*, pages 6978–6988, 2023.

[Zavrtanik *et al.*, 2021] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DrÆm – a discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, pages 8310–8319, 2021.

[Zhang *et al.*, 2023a] Hui Zhang, Zheng Wang, Zuxuan Wu, and Yu-Gang Jiang. Diffusionad: Norm-guided one-step denoising diffusion for anomaly detection. *arXiv*, 2023.

[Zhang *et al.*, 2023b] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *CVPR*, pages 3914–3923, 2023.

[Zhou *et al.*, 2022] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, page 2337–2348, sep 2022.

[Zhu *et al.*, 2023] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv*, 2023.

[Zou *et al.*, 2022] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. page 392–408. Springer-Verlag, 2022.