PromptAD: Learning Prompts with only Normal Samples for Few-Shot Anomaly Detection

Xiaofan Li[1], Zhizhong Zhang[1], Xin Tan[1,2,†], Chengwei Chen[2], Yanyun Qu[3], Yuan Xie[1,2], Lizhuang Ma[1]
[1]East China Normal University, Shanghai, China
[2]Chongqing Institute of East China Normal University, Chongqing, China
[3]The Navy Military Medical University, Shanghai, China
[4]Xiamen University, Fujian, China
{funzi}@stu.ecnu.edu.cn, {zzzhang, xtan, yxie, lzma}@cs.ecnu.edu.cn
{timchen}91@aliyun.com, {yyqu}@xmu.edu.cn

Abstract

The vision-language model has brought great improvement to few-shot industrial anomaly detection, which usually needs to design of hundreds of prompts through prompt engineering. For automated scenarios, we first use conventional prompt learning with many-class paradigm as the baseline to automatically learn prompts but found that it can not work well in one-class anomaly detection. To address the above problem, this paper proposes a one-class prompt learning method for few-shot anomaly detection, termed PromptAD. First, we propose semantic concatenation which can transpose normal prompts into anomaly prompts by concatenating normal prompts with anomaly suffixes, thus constructing a large number of negative samples used to guide prompt learning in one-class setting. Furthermore, to mitigate the training challenge caused by the absence of anomaly images, we introduce the concept of explicit anomaly margin, which is used to explicitly control the margin between normal prompt features and anomaly prompt features through a hyper-parameter. For image-level/pixel-level anomaly detection, PromptAD achieves first place in 11/12 few-shot settings on MVTec and VisA. Code is available at https://github.com/FuNz-0/PromptAD.git

1. Introduction

Anomaly detection (AD) [4, 35, 48] is a critical task in computer vision [22, 25, 28, 29], with widespread applications of defect detection in industry and medicine. This paper focuses on unsupervised industrial anomaly detection, which poses a challenge known as a one-class classification (OCC)

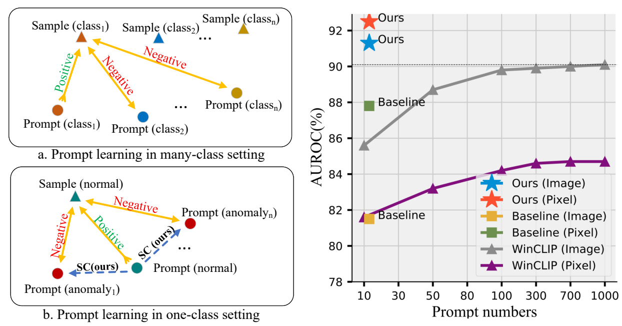


Figure 1. **Left:** Prompt learning under many-class and one-class settings. **Right:** The prompt-guided results of WinCLIP using different numbers of prompts, and the prompt-guided results of the baseline and our PromptAD under one-shot for prompt learning. All results are on the MVTec.

[37] setting. In this framework, only normal samples are available during training, but in the testing phase, the model is expected to identify anomalous samples. Since industrial anomaly detection typically customizes a model for various industrial production lines, the ability to rapidly train models with few samples holds significant promise for practical applications.

Due to the strong zero-shot ability of the foundation models [23, 32, 34], WinCLIP [19] was proposed as the first work utilizing the vision-language foundation model (*i.e.*, CLIP [33]) to enhance the model's anomaly detection performance in few-shot settings. To better leverage prompt guidance, WinCLIP introduces a prompt engineer strategy called "Prompt Ensemble" which combines a sufficient number of manually-designed prompts. For example, some manual prompts (*e.g.*, `a cropped photo of a []`, `a blurry photo of the []`, *etc.*) are combined together as the normal prompts. As shown in Figure 1 (right), with the number of prompts increasing, WinCLIP's perfor-

[†]Corresponding author.

mance improves, reaching a saturation point at around 1000 prompts. Other methods like SAA+ [5] and AnoVL [11] also employ prompt engineering to enhance model performance, which has become a rite of prompt-guided anomaly detection. Prompt engineering involves human intervention and requires careful design, which does not meet the automation requirements of industrial scenarios.

As illustrated in Fig.1 (left a.), prompt learning [55] aims to automatically learn prompts through contrastive learning [6, 16] for guiding image classification. The idea of prompt learning for anomaly detection is intriguing. However, as shown in Figure 1 (right), due to the one-class setting of anomaly detection, using the above prompt learning paradigm [55] as the baseline does not work well and is inferior to WinCLIP [19] with manual prompts on the image-level result. The main challenges are as follows: 1) prompt learning relies on contrastive learning, how to design prompts to complete the contrastive learning in the one-class setting? 2) With the absence of anomaly samples, how to control the marginal distance between normal prompts and anomaly prompts?

In this paper, we propose the one-class prompt learning with only normal samples for AD termed **PromptAD**. To solve the first challenge above, we propose semantic concatenation (SC). Intuitively, concatenating a prompt with antisense texts can transpose its semantics. According to this idea, as illustrated in Figure 1 (left b.), SC first designs a learnable normal prompt such as $[\mathbf{P}_1][\mathbf{P}_2]\dots[\mathbf{P}_{E_N}][obj.]$ for normal samples, and then manually concatenate various texts related to anomalies with the normal prompt such as $[\mathbf{P}_1][\mathbf{P}_2]\dots[\mathbf{P}_{E_N}][obj.][with][flaw]$ which is converted into an anomaly prompt and can be used as a negative prompt of normal sample during prompt learning. Due to the manually annotated anomalous texts are very limited. To expand the richness of anomaly information, SC also designs learnable anomaly prompts by concatenating a suffix of learnable tokens with a normal prompt, for instance $[\mathbf{P}_1][\mathbf{P}_2]\dots[\mathbf{P}_{E_N}][obj.][\mathbf{A}_1][\mathbf{A}_2]\dots[\mathbf{A}_{E_A}]$, where $[\mathbf{A}_i]$ is learnable token. The distribution of learnable anomaly prompts and manual anomaly prompts are aligned to ensure that the learnable anomaly prompts learn more correct anomaly information.

Furthermore, in anomaly detection, anomaly samples are unavailable, making it impossible to explicitly control the margin between normal and anomaly prompt features through contrastive loss. To address the second challenge, we propose the concept of Explicit Anomaly Margin (EAM), where a hyper-parameter is introduced to ensure that the distance between normal features and normal prompt features is smaller than the distance between normal features and anomaly prompt features. Thus ensuring a sufficient margin between normal prompts and anomaly prompts. Figure 1 (right) illustrates our great advantages, it can be seen that (compared with the WinCLIP [19] and Baseline [55]) PromptAD achieves 91.3%($\uparrow$1.2% and $\uparrow$9.8%)/92.5%($\uparrow$7.7% and $\uparrow$3.7%) image-level/pixel-level anomaly detection results with only 10$\sim$20 ($\downarrow$ $\sim$980 and $\downarrow$ 0) prompts.

To summarize, the main contributions of this paper are:

- We explore the feasibility of prompt learning in one-class anomaly detection, and propose a one-class prompt learning method termed **PromptAD**, which thoroughly beats conventional many-class prompt learning.
- Semantic concatenation (SC) is proposed, which can transpose the semantics of normal prompts by concatenating anomaly suffixes, so as to construct enough negative prompts for normal samples.
- Explicit anomaly margin (EAM) is proposed, which can explicitly control the distance between normal prompt features and anomaly prompt features through a hyper-parameter.
- For image-level/pixel-level anomaly detection, PromptAD achieves first place in 11/12 few-shot settings on MVTec [4] and VisA [57].

2. Related Work

Vision-Language Model. Leveraging contrastive learning [6, 16] and vision transformer [57], some vision-language models (VLM) [2, 20, 23, 33] have recently achieved great success. CLIP is one of the most commonly used VLMs, which is trained on web-scale image-text and shows strong zero-shot classification ability. The code of CLIP for LAION-400M [39] and LAION-5B [40] scale pre-training is open-scoured by OpenCLIP [18]. With the pre-trained CLIP and prompt engineer, huge leaps were made for some downstream tasks [7, 24, 47, 56]. Influenced by the success of prompt learning [21, 42] in Natural Language Processing (NLP), there has been a surge of prompt learning methods [13, 46, 54, 55] in recent times for few-shot image classification tasks. These methods aim to automatically learn better prompts through contrastive learning [6, 16] for guiding image classification based on CLIP.

Anomaly Detection. Most of the AD methods mainly focus on three paradigms: feature embedding paradigm, knowledge distillation paradigm, and reconstruction-based paradigm. The feature embedding paradigm [1, 9, 10, 30, 31, 35, 36, 49] extracts the patch features of the image through the neural network and then performs anomaly detection,. The knowledge distillation paradigm [3, 14, 38, 45, 45, 52] lets the student network only learn the knowledge of the normal samples of the teacher network, and complete anomaly detection through the difference between the teacher and the student. The reconstruction paradigm [15, 50, 51] hopes that the model can reconstruct the anomaly image into a normal image, and realize anomaly

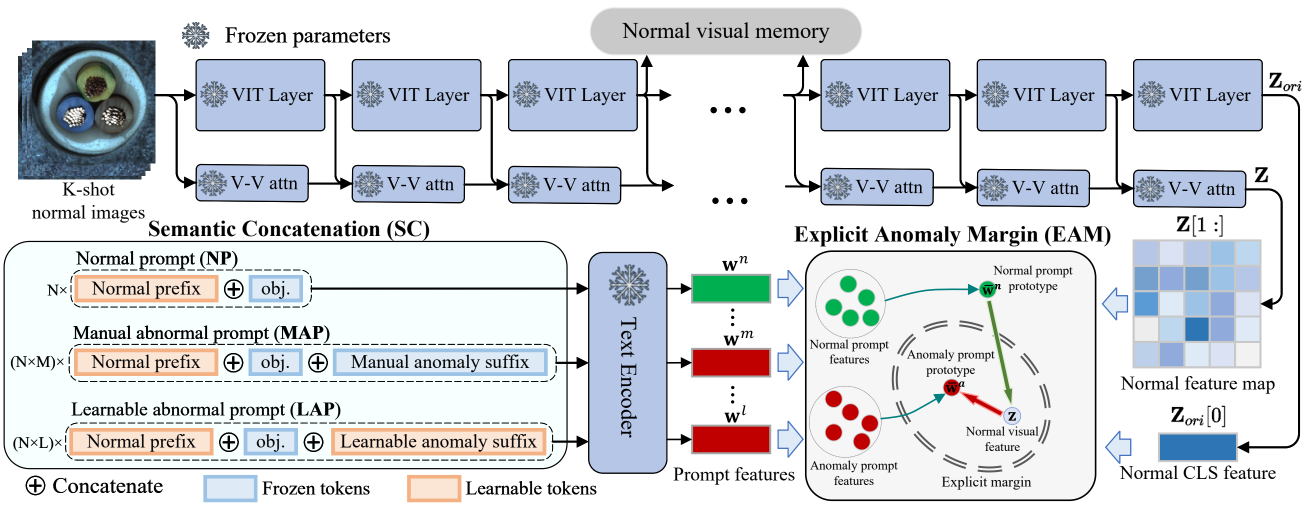


Figure 2. Illustration of PromptAD, which includes two novel modules: SC and EAM. The visual encoder has been transformed with v-v attention. The original branch is used to extract CLS feature, while the v-v attention branch is used to extract the feature map.

detection by the difference between the reconstructed image and the anomaly image.

Few-Shot Anomaly Detection. TDG [41] and RegAD [17] are the first to explore few-shot anomaly detection methods, and PatchCore [35] and DifferNet [36] also demonstrated the performance in few-shot settings. WinCLIP [19] and RWDA [43] introduce the CLIP model to anomaly detection and greatly improve the performance in the few-shot setting. The latest FastRecon [12] reconstructs anomaly features by regression with distribution regularization and achieves excellent performance.

3. Preliminaries

3.1. CLIP and Prompt Learning

Contrastive Language Image Pre-training termed CLIP [33] is a large-scale vision-language model which is famous for its zero-shot classification ability. Specifically, giving an unknown image $\mathbf{i}$, and K text-prompts $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_K\}$, CLIP can predict the distribution of $\mathbf{i}$ belonging to these K text-prompts:

$$p(\mathbf{y}|\mathbf{i}) = \frac{\exp < f(\mathbf{i}), g(\mathbf{s}_y)/\tau >}{\sum_{i=1}^{K} \exp < f(\mathbf{i}), g(\mathbf{s}_i)/\tau >}, \tag{1}$$

where $f(\cdot)$ and $g(\cdot)$ are visual and text encoder respectively. $< \cdot, \cdot >$ represents cosine similarity, τ is the temperature hyper-parameter. The initial text prompt used for CLIP zero-shot classification is still simple, such as `a photo of [class]`, *etc.*, slightly better than directly using the name of the class as the prompt.

Prompt Learning. Inspired by the success of prompt learning in natural language processing (NLP) [21, 42], CoOp [55] introduces this paradigm into few-shot classification, aiming to automatically learn efficient prompts for CLIP. Specifically, the prompt used in CoOp is not the frozen text description, but a set of trainable parameters:

$$\mathbf{s}_k = [\mathbf{P}_1][\mathbf{P}_2]\dots[\mathbf{P}_{E_P}][class_k], \tag{2}$$

where $[\mathbf{P}_1][\mathbf{P}_2]\dots[\mathbf{P}_{E_P}]$ are trainable tokens and $[class_k]$ is k-th class name which is not trainable. Prompt learning aims to automatically train effective prompts to improve CLIP performance on downstream classification tasks.

3.2. CLIP Surgery

As a classification model, CLIP is far less adaptive in prompt-guided image localization tasks without fine-tuning. To find out why CLIP fails to image localization tasks, some CLIP explainable works [27, 53] analyze the mechanism that how CLIP extracts visual features. These works observed that the global feature extraction of Q-K self-attention [44] affects the localization ability of CLIP, which is as follows:

$$Attn(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\mathbf{Q} \cdot \mathbf{K}^{\mathrm{T}} \cdot scale) \cdot \mathbf{V}. \tag{3}$$

To this end, CLIP-Surgery [27] proposes a V-V attention mechanism to enhance the model's attention to local features without destroying the original structure. As shown in Figure 2, the feature extraction process is described as follows:

$$\mathbf{Z}_{ori}^{l-1} = [\mathbf{t}_{cls}; \mathbf{t}_1; \mathbf{t}_2, ...; \mathbf{t}_T], \tag{4}$$

$$\mathbf{Z}^{l-1} = [\mathbf{t}'_{cls}; \mathbf{t}'_1; \mathbf{t}'_2, ...; \mathbf{t}'_T], \tag{5}$$

$$[\mathbf{Q}^l, \mathbf{K}^l, \mathbf{V}^l] = QKV\_Proj.^l(\mathbf{Z}_{ori}^{l-1}), \tag{6}$$

$$\mathbf{Z}^l = Proj.^l(Attn(\mathbf{V}^l, \mathbf{V}^l, \mathbf{V}^l)) + \mathbf{Z}^{l-1}, \tag{7}$$

where $\mathbf{Z}_{ori}^{l-1}$ denotes the $(l-1)$-th layer output of the original CLIP visual encoder and $\mathbf{Z}^{(l-1)}$ denotes the local-aware output of layer $l-1$, $QKV\_Proj.^l$ and $Proj^l$ denote the QKV projection and output projection whose parameters are initialized by the visual encoder parameters of the original CLIP. The final original outputs and local-aware outputs are $\mathbf{Z}_{ori}$ and $\mathbf{Z}$, the CLS feature $\mathbf{Z}_{ori}[0] \in \mathbb{R}^d$ is used for image-level anomaly detection and the local feature map $\mathbf{Z}[1:] \in \mathbb{R}^{T\times d}$ is used for pixel-level anomaly detection. In this paper, we use modified CLIP as the backbone and term it VV-CLIP.

4. Methodology

4.1. Overview

An overview of our proposed PromptAD is illustrated in Figure 2. PromptAD is built on VV-CLIP whose visual encoder is used to extract global and local features. The proposed semantic concatenation (SC) is used to design prompts. Specifically, N learnable normal prefixes and the objective name are concatenated to get normal prompts (NPs), then N normal prompts are concatenated with M manual anomaly suffixes and L learnable anomaly suffixes respectively to obtain $N \times M$ manual anomaly prompts (MAPs) and $N \times L$ learnable anomaly prompts (LAPs). The visual features and prompt features are used to complete prompt learning by contrastive loss and the proposed explicit anomaly margin (EMA) loss. EMA can control the explicit margin between the normal prompt features and anomaly prompt features through a hyper-parameter. Finally, the prompts obtained by prompt learning are used for prompt-guided anomaly detection (PAD).

In addition to PAD, referring to WinCLIP+ [19], we also introduce vision-guided anomaly detection (VAD). Specifically, as shown in Figure 2, during training, the i-th layer features (without CLS feature) output by the visual encoder are stored as normal visual memory which is denoted as $\mathbf{R}$. In the testing phase, the ith layer feature map $\mathbf{F} \in \mathbb{R}^{h\times w\times d}$ of a query image is compared with $\mathbf{R}$ to obtain the anomaly score map $\mathbf{M} \in [1, 0]^{h\times w}$:

$$\mathbf{M}_{ij} = \min_{\mathbf{r}\in\mathbf{R}} \frac{1}{2}(1- < \mathbf{F}_{ij}, \mathbf{r} >). \quad (8)$$

In practice, we use the intermediate features of two layers as memory to get two score maps for each query image and then average the two score maps to get the final vision-guided score map $\mathbf{M}_v$.

4.2. Semantic Concatenation

Only normal samples are obtainable during anomaly detection training, which leads to no negative samples for guiding prompt learning and thus impairs its effect. We found that the semantics of prompts can be changed by concatenating. For example, `a photo of cable` has normal semantics, and after concatenating it with a suffix, `a photo of cable with flaw` is converted into anomaly semantics. In this way, we propose semantic concatenation (SC) which can transpose normal prompts to anomaly prompts by concatenating normal prompts with anomaly suffixes, so as to construct sufficient contrast prompts based on learnable normal prompts. Specifically, following the format of CoOp [55], the learnable normal prompt (NP) is designed as follows:

$$\mathbf{s}^n = [\mathbf{P}_1][\mathbf{P}_2]\dots[\mathbf{P}_{E_N}][obj.], \quad (9)$$

where E_N denotes the length of the learnable normal prefix and $[obj.]$ represents the name of the object being detected. The learnable normal prompt can be transposed to an anomaly prompt after concatenating with the anomaly suffixes. In particular, we generated anomaly suffixes from the anomaly labels of the datasets [4, 57], such as `[] with color stain`, `[] with crack`, *etc.*, and then concatenate these texts with the NP to obtain the manual anomaly prompt (MAP):

$$\mathbf{s}^m = [\mathbf{P}_1][\mathbf{P}_2]\dots[\mathbf{P}_{E_N}][obj.][with][color][stain], \quad (10)$$

where the prefix is a trainable NP and the suffix is a manual anomaly text. In addition, we combine NP with a learnable token suffix to design the learnable anomaly prompt (LAP):

$$\mathbf{s}^l = [\mathbf{P}_1][\mathbf{P}_2]\dots[\mathbf{P}_{E_N}][obj.][\mathbf{A}_1]\dots[\mathbf{A}_{E_A}], \quad (11)$$

where E_A denotes the length of learnable anomaly suffix. It should be noted that the parameters of prompts concatenated by the same normal prefix or anomaly suffix are shared. During training, NPs move close to normal visual features, while MAPs and LAPs move away from normal visual features. The training loss for prompt learning is consistent with the CLIP training loss as follows:

$$\mathcal{L}_{clip} = \mathbb{E}_{\mathbf{z}}\left[-log\frac{\exp(<\mathbf{z}, \bar{\mathbf{w}}^n/\tau>)}{\exp < \mathbf{z}, \bar{\mathbf{w}}^n/\tau > + \sum\limits_{\mathbf{w}\in\mathcal{W}} \exp < \mathbf{z}, \mathbf{w}/\tau >}\right], \quad (12)$$

where $\mathbf{z}$ denotes normal visual feature, $\bar{\mathbf{w}}^n = \frac{\sum_{i=1}^N g(\mathbf{s}_i^n)}{N}$ is the prototype of normal prompt features, $\mathcal{W} = \{g(\mathbf{s})|\mathbf{s} \in \text{MAPs} \cup \text{LAPs}\}$ is a set containing all anomaly prompt features. Since more negative samples can produce a better contrastive learning effect [16], each anomaly prompt feature is compared with the visual feature

Remark. In the one-class anomaly detection, conventional prompt learning can only design learnable normal prompts, which is not conducive to the effect of contrastive loss. The proposed semantic concatenation can transform the semantics of normal prompts into anomaly semantics with shared parameters, which can make normal samples contrast with the semantic transposes (anomaly prompts).

4.3. Explicit Anomaly Margin

Due to the lack of anomaly visual samples in the training, the MAPs and LAPs can only take normal visual features as negative samples for contrast and lack an explicit margin between the normal and anomaly prompts. Therefore, we propose the explicit anomaly margin (EAM) for AD prompt learning, which can control the margin between normal prompt features and anomaly prompt features. EAM is actually a regularization loss implemented via a margin hyper-parameter, which is defined as:

$$\mathcal{L}_{ema} = \mathbb{E}_{\mathbf{z}}\left[\max\left(0, d(\frac{\mathbf{z}}{\|\mathbf{z}\|_2}, \frac{\bar{\mathbf{w}}^n}{\|\bar{\mathbf{w}}^n\|_2}) - d(\frac{\mathbf{z}}{\|\mathbf{z}\|_2}, \frac{\bar{\mathbf{w}}^a}{\|\bar{\mathbf{w}}^a\|_2})\right)\right], \tag{13}$$

where $d(\cdot,\cdot)$ represents euclidean distance, and $\bar{\mathbf{w}}^a$ is the prototype of all anomaly prompt features:

$$\bar{\mathbf{w}}^a = \frac{\sum_{i=1}^{N\times M} g(\mathbf{s}_i^m) + \sum_{i=1}^{N\times L} g(\mathbf{s}_i^l)}{N\times M + N\times L}. \tag{14}$$

In CLIP, the final features are all projected onto the unit hyper-sphere, thus the features in $\mathcal{L}_{ema}$ are also normalized, and the margin is fixed to zero. Compared to contrastive loss ($\mathcal{L}_{clip}$), EMA loss guarantees a larger distance between normal samples and the anomaly prototype than between normal samples and the normal prototype, resulting in an explicit discrimination between normal and anomaly prototypes.

In addition, since MAPs contain sufficient anomaly information while LAPs are initialized without any semantic guidance, aligning them helps LAPs to mimic the distribution of MAPs. Specifically, we align the means of the two distributions using the squared l_2 norm:

$$\mathcal{L}_{align} = \lambda \cdot \left\| \frac{\bar{\mathbf{w}}^m}{\|\bar{\mathbf{w}}^m\|_2} - \frac{\bar{\mathbf{w}}^l}{\|\bar{\mathbf{w}}^l\|_2} \right\|_2^2, \tag{15}$$

where $\bar{\mathbf{w}}^m$ and $\bar{\mathbf{w}}^l$ are the feature means of MAPs and LAPs, respectively, and λ is a hyper-parameter controlling the alignment degree of MAPs and LAPs.

4.4. Anomaly Detection

In the testing phase, $\bar{\mathbf{w}}^n$ is used as the normal prototype and $\bar{\mathbf{w}}^a$ is used as the anomaly prototype to complete prompt-guided anomaly detection. The image-level score $\mathbf{S}_t \in [0,1]$ and pixel-level score map $\mathbf{M}_t \in [0,1]^{h\times w}$ are calculated through:

$$score = \frac{\exp < \mathbf{z}_t, \bar{\mathbf{w}}^n/\tau >}{\exp < \mathbf{z}_t, \bar{\mathbf{w}}^n/\tau > + \exp < \mathbf{z}_t, \bar{\mathbf{w}}^a/\tau >}, \tag{16}$$

where $\mathbf{z}_t$ is a global/local image feature for image-level/pixel-level anomaly detection.

Finally, vision-guided $\mathbf{M}_v$ and prompt-guided $\mathbf{M}_t$ are fused to obtain the pixel-level anomaly score map, and the maximum value of $\mathbf{M}_v$ and $\mathbf{S}_t$ are fused to obtain the image-level anomaly score:

$$\mathbf{M}_{pix} = 1.0/(1.0/\mathbf{M}_v + 1.0/\mathbf{M}_t), \tag{17}$$

$$\mathbf{S}_{img} = 1.0/(1.0/\max_{ij} \mathbf{M}_v + 1.0/\mathbf{S}_t), \tag{18}$$

where the fusion method we use is harmonic mean, which is more sensitive to smaller values [19].

5. Experiments

We complete the comparison experiments between PromptAD and the latest methods under 1, 2, and 4-shot settings, which include both image-level and pixel-level results. In addition, we also compare the many-shot and full-shot methods to show the powerful few-shot performance of PromptAD. Finally, we conduct ablation experiments to verify the improvement of prompt learning by the proposed SC and EAM, and show the impact of different CLIP transformation methods [27, 53] and hyper-parameters.

Dataset. In this paper, the benchmarks we use are MVTec [4] and VisA [57]. Both benchmarks contain multiple subsets with only one object per subset. MVTec contains 15 objects with $700^2 - 900^2$ pixels per image, and VisA contains 12 objects with roughly $1.5\text{K} \times 1\text{K}$ pixels per image. Anomaly detection is a one-class task, so the training set contains only normal samples, while the test set contains normal samples and anomaly samples with image-level and pixel-level annotations. In addition, the anomaly category present for each object is also annotated.

Evaluation metrics. We follow the literature [4] in reporting the Area Under the Receiver Operation Characteristic (AUROC) for both image-level and pixel-level anomaly detection.

Implementation details. We used the OpenCLIP [18] implementation of CLIP and its pre-trained parameters, in addition to the default values of the hyper-parameter τ. Referring to WinCLIP [19], we used LAION-400M [39] based CLIP with ViT-B/16+.

5.1. Image-level Comparison Results

The Image-level comparative experimental results of PromptAD and current methods are recorded in Table 1, where SPADE [9], PaDiM [10], and PatchCore [35] are the reformulations of traditional full-shot methods in the few-shot settings. It can be seen that the Image-level AD performance of these methods is very limited. Both WinCLIP+ [19] and RWDA [43] introduce CLIP [33], which greatly improves the performance of Image-level AD under few-shot settings. Compared with the above methods, PromptAD achieves significant improvement in three settings of the two benchmarks. Compared with WinCLIP+

Method	Public	MVTec			VisA		
		1-shot	2-shot	4-shot	1-shot	2-shot	4-shot
SPADE [9]	arXiv'2020	81.0±2.0	82.9±2.6	84.8±2.5	79.5±4.0	80.7±5.0	81.7±3.4
PaDiM [10]	ICPR'2020	76.6±3.1	78.9±3.1	80.4±2.4	62.8±5.4	67.4±5.1	72.8±2.9
PatchCore [35]	CVPR'2022	83.4±3.0	86.3±3.3	88.8±2.6	79.9±2.9	81.6±4.0	85.3±2.1
WinCLIP+†[19]	CVPR'2023	93.1±2.0	94.4±1.3	95.2±1.3	83.8±4.0	84.6±2.4	87.3±1.8
RWDA†[43]	BMVC'2023	93.3±0.5	94.0±0.7	94.5±0.7	83.4±1.7	85.6±1.4	86.6±0.9
FastRcon [12]	ICCV'2023	-	91.0	94.2	-	-	-
PromptAD†	-	**94.6**±1.7	**95.7**±1.5	**96.6**±0.9	**86.9**±2.3	**88.3**±2.0	**89.1**±1.7

Table 1. Comparison of image-level anomaly detection in AUROC on MVTec and VisA benchmarks. The best and second-best results are respectively marked in bold and underlined. † indicates CLIP-based methods.

Method	Public	Setting	image	pixel
PromptAD	-	1-shot	94.6	95.9
PromptAD	-	4-shot	96.6	96.5
DiffNet [36]	WACV'2021	16-shot	87.3	-
TDG [41]	ICCV'2021	10-shot	78.0	-
RegAD [17]	ECCV2022	8-shot	91.2	96.7
FastRecon [12]	ICCV'2023	8-shot	95.2	97.3
MKD [38]	CVPR'2021	full-shot	87.8	90.7
P-SVDD [49]	ACCV'2021	full-shot	95.2	96.0
PatchCore [35]	CVPR'2022	full-shot	99.1	98.1
SimpleNet [30]	CVPR'2023	full-shot	99.6	98.1

Table 2. Comparison with exiting many-shot methods in AUROC (image and pixel level) on MVTec. Results below our 1-shot are marked in red, and those below our 4-shot are marked in blue.

and RWDA, PromptAD achieves 1.3%, 1.3%, and 1.4% (2.9%, 2.7%, 1.8%) improvement under the 1, 2, and 4-shot Settings of MVTec (and VisA), respectively. In addition, PromptAD uses a smaller number of prompts than WinCLIP+ and RWDA.

5.2. Pixel-level Comparison Results

The pixel-level comparative experimental results are recorded in Table 4. It can be seen that the CLIP-based method (WinCLIP+ [19]) and other methods perform comparably on pixel-level AD, and the improvement brought by the introduction of CLIP [33] is not as obvious as that on image-level AD. PromptAD achieves the best place on MVTec/VisA in the 1-shot and 2-shot settings, which are 0.7%/0.3% and 0.2%/0.3% higher than WinCLIP+, respectively. In the 4-shot setting, while PromptAD ranks first on VisA, it takes second place on MVTec, narrowly outperformed by FastRecon [12] with a 0.5% margin.

The quantitative results of anomaly localization are shown in Figure 3. Compared with PatchCore [35] and WinCLIP+ [19], PromptAD has a better anomaly localization capability for both objects and textures in the 1-shot setting. In addition, PromptAD can also locate some very small anomaly areas very accurately.

5.3. Compared With Many-shot Methods

In Table 2, the comparison results of PromptAD under few-shot settings with other methods under many-shot/full-shot settings are recorded. It can be seen that compared with some methods under many-shot settings, PromptAD achieves better image-level results, and the pixel-level results are also competitive, which fully verifies the strong ability of PromptAD in the few-shot settings. In addition, PromptAD is superior to the early full-shot AD methods, MKD [38] and P-SVDD [49], but there is still a certain gap between PromptAD and the latest full-shot AD methods, PatchCore [35] and SimpleNet [30].

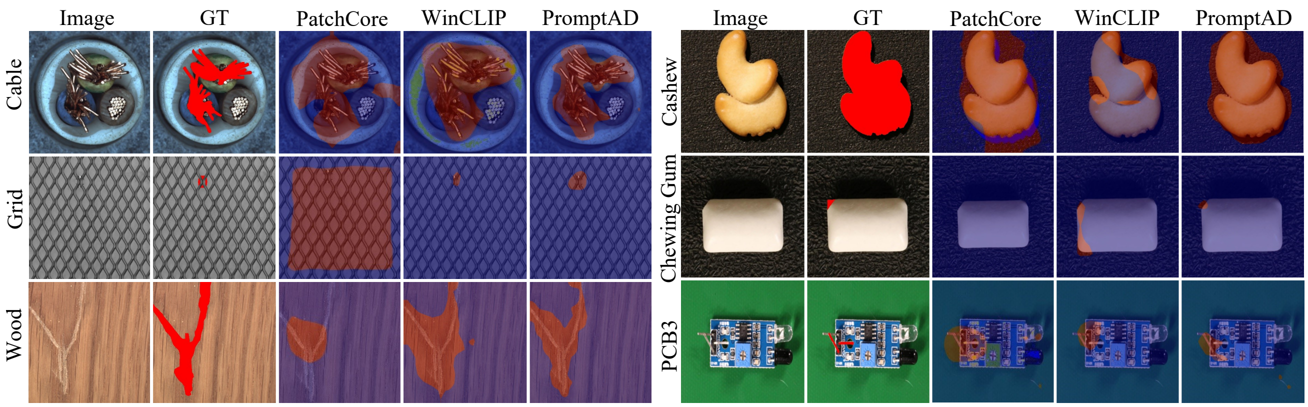
PAD		VAD	MVTec		VisA	
SC	EAM		image	pixel	image	pixel
✗	✗	✗	81.5	87.8	72.6	85.5
✓	✗	✗	90.4	91.7	81.3	90.5
✓	✓	✗	91.3	92.5	83.2	91.8
✗	✗	✓	85.1	93.2	82.7	95.2
✓	✓	✓	**94.6**	**95.9**	**86.9**	**96.7**

Table 3. Image-level/pixel-level results (AUROC) of ablation study under 1-shot setting. PAD and VAD are prompt-guided and vision-guided anomaly detection, respectively, SC is semantic concatenation, and EAM is explicit anomaly margin.

5.4. Ablation Study

We verify the impact of different modules of different proposed methods on the overall performance of PromptAD under 1-shot setting on MVTec [4] and VisA [57]. These include semantic concatenation (SC) and explicit anomaly margin (EMA). Meanwhile, we also verified the effect of vision-guided anomaly detection (VAD). Results of the ablation study are recorded in Table 3.

Semantic Concatenation (SC). The number of negative samples plays a crucial role in contrastive learning [6, 16]. Without the proposed SC, the conventional prompt learning paradigm [55] loses negative prompts for contrast, so the effect of prompt learning will be greatly reduced. As shown in Table 3, there is a significant drop in image and pixel level results on both MVTec [4] and VisA [57] when SC is not used. After using SC, the image-level/pixel-level results on MVTec (and VisA) are improved by 8.9%/3.9% (8.9%/5.0%), which indicates that SC can greatly improve the applicability of prompt learning in anomaly detection.



(a) MVTec (1-shot)　　(b) VisA (1-shot)

Figure 3. Qualitative comparison results of 1-shot pixel-level anomaly detection on MVTec [4] and VisA [57].

Method	Public	MVTec			VisA		
		1-shot	2-shot	4-shot	1-shot	2-shot	4-shot
SPADE [9]	arXiv'2020	91.2±0.4	92.0±0.3	92.7±0.3	95.6±0.4	96.2±0.4	96.6±0.3
PaDiM [10]	ICPR'2020	89.3±0.9	91.3±0.7	92.6±0.7	89.9±0.8	92.0±0.7	93.2±0.5
PatchCore [35]	CVPR'2022	92.0±1.0	93.3±0.6	94.3±0.5	95.4±0.6	96.1±0.5	96.8±0.3
WinCLIP+† [19]	CVPR'2023	<u>95.2±0.5</u>	<u>96.0±0.3</u>	96.2±0.3	<u>96.4±0.4</u>	<u>96.8±0.3</u>	<u>97.2±0.2</u>
FastRecon [12]	ICCV'2023	-	95.9	**97.0**	-	-	-
PromptAD†	-	**95.9**±0.5	**96.2**±0.3	<u>96.5±0.2</u>	**96.7**±0.4	**97.1**±0.3	**97.4**±0.3

Table 4. Comparison of pixel-level anomaly detection in AUROC on MVTec and VisA benchmarks. The best and second-best results are respectively marked in **bold** and underlined. † indicates CLIP-based methods.

Explicit Anomaly Margin (EAM). Since anomaly samples are absent during the training phase, it is hard to establish an explicit margin between the features of normal and anomaly prompts. EAM uses a hyper-parameter to control the margin between normal and anomaly prompt features, which can make up for the lack of contrastive loss. Table 3 shows that after using EAM, the image-level/pixel-level results on MVTec (and VisA) are improved by 0.9%/0.8% (1.9%/0.7%), respectively.

Vision-guided Anomaly Detection (VAD). PAD introduces more high-level semantic information but ignores many local details, which is not conducive to pixel-level anomaly detection. On the contrary, VAD using normal feature memory focuses on more local detail information. In Table 3, PAD has better image-level results, while VAD has better pixel-level results, and the two have a good complementarity. Under the 1-shot setting, the results of PAD and VAD are fused by harmonic mean, and 94.6%/95.9% (86.8%/96.7%) image-level/pixel-level results are achieved on MVTec (and VisA).

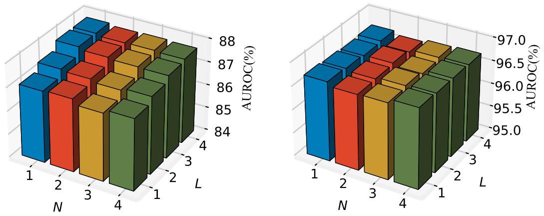
5.5. Results of Different CLIP Transformations

Due to the inability of CLIP to directly complete prompt-guided localization tasks, some works have explored the transformations of CLIP [27, 53]. Table 5 records the results of different CLIP transformations under pixel-level anomaly detection, where about 1000 prompts are used in 0-shot setting and our prompt learning method is used in 1-shot setting. MaskCLIP [53] drops the QKV attention and leaves only $V\_Proj.$ and $Proj.$, and then embed local features after each layer of the visual encoder as in VV-CLIP. CLIP+Linear [8] adds a learnable linear layer to the visual encoder after each block to align the local features with prompt features.

Method	MVTec	VisA
CLIP [33]	22.5	24.6
CLIP [33] + ours	79.9	80.4
CLIP+Linear [8] + ours	79.4	77.2
MaskCLIP [53]	85.5	80.5
MaskCLIP [53] + ours	<u>91.6</u>	<u>91.2</u>
VV-CLIP [27]	86.7	82.9
VV-CLIP [27] + ours	**92.5**	**91.8**

Table 5. Pixel-level results (AUROC) of using different CLIP transformations on MVTec and VisA under 0-shot/1-shot settings.

As shown in Table 5, the results of the original CLIP under 0-shot are 22.5%/24.6% on MVTec/Visa, which is lower than the random prediction (50.0%). This is caused by the opposite visual activation [26, 27] of CLIP. After the transformation of attention, MaskCLIP [53] and VV-CLIP [27] achieve a huge improvement of 63.0%/55.9% and 64.2%/58.3% on MVTec/Visa, respectively. The improvement of VV-CLIP is more obvious than that of MaskCLIP.



(a) Image-level results. (b) Pixel-level results.

Figure 4. Image-level/pixel-level results on VisA [57] in 1-shot setting using different N and L.

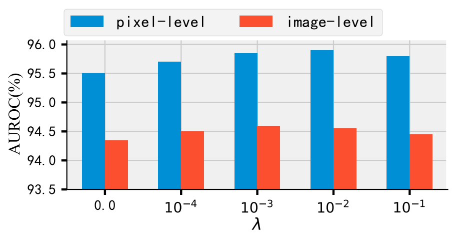


Figure 5. Image-level/pixel-level results on MVTec [4] in the 1-shot setting using different hyper-parameter λ.

We speculate that this is because VV-attention retains a certain information interaction while focusing on local information, while MaskCLIP completely removes attention.

After using our method, the pixel-level results of MaskCLIP and VV-CLIP are increased by 6.1%/10.7%, and 5.8%/9.8% on MVTec/Visa, respectively. Furthermore, it is worth noting that prompt learning also leads to a significant 57.4%/55.8% improvement in pixel-level results of the original CLIP. However, when prompt learning is added with learnable linear layers, the effect decreases, which may be because there is mutual interference between prompt learning and the training of linear layers.

5.6. Hyper-parameter Analysis

We complete the effect of N, L and λ on PromptAD. λ is the hyper-parameter of the loss $\mathcal{L}_{algin}$, which controls the degree of alignment between MAPs and LAPs feature distributions. N is the number of NPs. L is the number of anomaly prompt suffixes, and $N \times L$ is the number of LAPs.

Figure 4 illustrates the effect of, N and L on PromptAD. In Image-level results, N does not have a great influence, and there is no significant difference between $N = 1$ and $N = 4$. While, L has a significant influence, and larger L can lead to higher results. In pixel-level results, the effects of both N and L are relatively small, and larger L slightly improves the results. Figure 5 records the image-level/pixel-level results with different λ. It can be seen that the results are worse when the λ is equal to 0 or larger. This indicates that the distributions of MAPs and LAPs need to be aligned, but not over-aligned, which will reduce the diversity of the anomaly prompts and thus reduce the model's perception of anomaly image features.

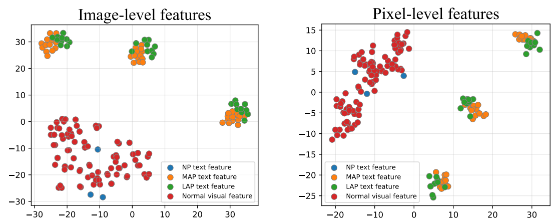


Figure 6. Feature visualization results using T-SNE in 1-shot setting. The feature used is "cable" from the MVTec [4].

5.7. Visualization Results

To quantify the effect of PromptAD, we visualize the visual and textual features after L_2 normalization. Specifically, we visualize 3 NPs, 3×13 MAPs, and 3×10 LAPs as well as 100 image-level/pixel-level normal visual features. Figure 6 shows the visualization results, it can be seen that there is very clear discrimination between normal prompt features and anomaly prompt features, and the overlap between normal prompt features and normal visual features is very high, which intuitively verifies the effectiveness of PromptAD. In addition, it's worth noting that the 3 normal prompt features do not collapse to one point, but fit the overall distribution of normal visual features as much as possible.

6. Conclusion

In this paper, we propose a novel anomaly detection method termed PromptAD which automatically learns prompts with only normal samples in the few-shot anomaly detection scenario. First, in order to cope with the challenge under the one-class task, we propose semantic concatenation to construct enough anomaly prompts through concatenating normal prompts and anomaly suffixes to guide prompt learning. Second, we propose the explicit anomaly margin loss, which explicitly determines the margin between normal prompt features and anomaly prompt features through a hyper-parameter. Finally, for image-level/pixel-level anomaly detection, PromptAD achieves first place in 11/12 few-shot tasks.

Acknowledgments This work is supported by the National Natural Science Foundation of China No. 62176092, 62222602, U23A20343, 62106075, 62302167, Shanghai Sailing Program (23YF1410500), Natural Science Foundation of Shanghai (23ZR1420400), Science and Technology Commission of Shanghai No.21511100700, Natural Science Foundation of Chongqing, China (CSTB2023NSCQ-JQX0007, CSTB2023NSCQ-MSX0137), CCF-Tencent Rhino-Bird Young Faculty Open Research Fund (RAGR20230121), China Postdoctoral Science Foundation funded project (2023M734270), and Development Project of Ministry of Industry and Information Technology (ZTZB.23-990-016).

References

[1] Jaehyeok Bae, Jae-Han Lee, and Seyun Kim. Pni: industrial anomaly detection using position and neighborhood information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6373–6383, 2023.

[2] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

[3] Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. *arXiv preprint arXiv:2303.14535*, 2023.

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

[5] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] Wentao Chen, Chenyang Si, Zhang Zhang, Liang Wang, Zilei Wang, and Tieniu Tan. Semantic prompt for few-shot image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23581–23591, 2023.

[8] Xuhai Chen, Yue Han, and Jiangning Zhang. A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2023.

[9] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020.

[10] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021.

[11] Hanqiu Deng, Zhaoxiang Zhang, Jinan Bao, and Xingyu Li. Anovl: Adapting vision-language models for unified zero-shot anomaly localization. *arXiv preprint arXiv:2308.15939*, 2023.

[12] Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17481–17490, 2023.

[13] 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. *International Journal of Computer Vision*, pages 1–15, 2023.

[14] Zhihao Gu, Liang Liu, Xu Chen, Ran Yi, Jiangning Zhang, Yabiao Wang, Chengjie Wang, Annan Shu, Guannan Jiang, and Lizhuang Ma. Remembering normality: Memory-guided knowledge distillation for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16401–16409, 2023.

[15] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6447–6458, 2023.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. Computer Vision Foundation / IEEE, 2020.

[17] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael W. Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *Computer Vision - ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIV*, pages 303–319. Springer, 2022.

[18] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021.

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

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[21] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8: 423–438, 2020.

[22] Xia Kong, Zuodong Gao, Xiaofan Li, Ming Hong, Jun Liu, Chengjie Wang, Yuan Xie, and Yanyun Qu. En-compactness: Self-distillation embedding & contrastive generation for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9306–9315, 2022.

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[24] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1405–1413, 2023.

[25] Xiaofan Li, Yachao Zhang, Shiran Bian, Yanyun Qu, Yuan Xie, Zhongchao Shi, and Jianping Fan. Vs-boost: Boosting visual-semantic association for generalized zero-shot learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 1107–1115. ijcai.org, 2023.

[26] Yi Li, Hualiang Wang, Yiqun Duan, Hang Xu, and Xiaomeng Li. Exploring visual interpretability for contrastive language-image pre-training. *arXiv preprint arXiv:2209.07046*, 2022.

[27] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023.

[28] Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. Deep dual consecutive network for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 525–534, 2021.

[29] Zhenguang Liu, Runyang Feng, Haoming Chen, Shuang Wu, Yixing Gao, Yunjun Gao, and Xiang Wang. Temporal feature alignment and mutual information maximization for video-based human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11006–11016, 2022.

[30] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411, 2023.

[31] Declan McIntosh and Alexandra Branzan Albu. Inter-realization channels: Unsupervised anomaly detection beyond one-class classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6285–6295, 2023.

[32] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

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

[36] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 1906–1915. IEEE, 2021.

[37] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

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

[39] 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. *arXiv preprint arXiv:2111.02114*, 2021.

[40] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[41] Shelly Sheynin, Sagie Benaim, and Lior Wolf. A hierarchical transformation-discriminating generative model for few shot anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8475–8484. IEEE, 2021.

[42] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

[43] Masato Tamura. Random word data augmentation with clip for zero-shot anomaly detection. *arXiv preprint arXiv:2308.11119*, 2023.

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

[45] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for anomaly detection. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 306. BMVA Press, 2021.

[46] Runqi Wang, Hao Zheng, Xiaoyue Duan, Jianzhuang Liu, Yuning Lu, Tian Wang, Songcen Xu, and Baochang Zhang. Few-shot learning with visual distribution calibration and cross-modal distribution alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23445–23454, 2023.

[47] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 11686–11695, 2022.

[48] Yao Wu, Mingwei Xing, Yachao Zhang, Yuan Xie, Zongze Wu, and Yanyun Qu. Perturbed progressive learning for semisupervised defect segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[49] Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian conference on computer vision*, 2020.

[50] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021.

[51] Xinyi Zhang, Naiqi Li, Jiawei Li, Tao Dai, Yong Jiang, and Shu-Tao Xia. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6782–6791, 2023.

[52] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3914–3923, 2023.

[53] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022.

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

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

[56] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023.

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