FE-CLIP: Frequency Enhanced CLIP Model for Zero-Shot Anomaly Detection and Segmentation

Tao Gong [1, 2, 3], Qi Chu [1, 2, 3*], Bin Liu[1, 2, 3], Wei Zhou[4], Nenghai Yu[1, 2, 3]
[1]School of Cyber Science and Technology, University of Science and Technology of China
[2]Anhui Province Key Laboratory of Digital Security
[3]the CCCD Key Lab of Ministry of Culture and Tourism
[4]Ling Yang Industrial Internet Co., Ltd.
{tgong, qchu, flowice, ynh}@ustc.edu.cn, wezhou8@iflytek.com

Abstract

Zero-shot anomaly detection (ZSAD) requires detection models trained using auxiliary data to detect anomalies without any training sample in a target dataset. It is challenging since the models need to generalize to anomalies across different domains. Recently, CLIP-based anomaly detection methods, such as WinCLIP and AnomalyCLIP, have demonstrated superior performance in the ZSAD task, due to the strong zero-shot recognition of the CLIP model. However, they overlook the utilization of frequency information of images. In this paper, we find that frequency information could benefit the ZSAD task, since some properties of the anomaly area, such as appearance defects, can also be reflected based on its frequency information. To this end, We propose Frequency Enhanced CLIP (FE-CLIP), taking advantage of two different but complementary frequency-aware clues, (1) Frequency-aware Feature Extraction adapter, and (2) Local Frequency Statistics adapter, in the visual encoder of CLIP, to deeply mine frequency information for the ZSAD task. We apply DCT as the frequency-domain transformation. Through comprehensive experiments, we show that the proposed FE-CLIP has good generalization across different domains and achieves superior zero-shot performance of detecting and segmenting anomalies in 10 datasets of highly diverse class semantics from various defect inspections and medical domains. Besides, FE-CLIP also achieves superior performance under the few-normal-shot anomaly detection settings.

1. Introduction

Anomaly Detection aims to predict an image or a pixel as normal or anomalous, and has been widely applied in various applications, such as industrial defect inspection [4, 5, 7, 16, 28, 35, 37, 47, 51] and medical image analysis [11, 27, 31, 42, 43]. Existing AD approaches typically assume that training examples in a target application domain are available for learning the detection models. However, this assumption may not hold in various scenarios, such as i) when accessing training data violates data privacy policies, or ii) when the target domain does not have relevant training data. Therefore, Zero-shot anomaly detection (ZSAD) is an emerging task for AD in such scenarios, to which the AD methods mentioned above are not viable, as it requires detection models to detect anomalies without any training sample in a target dataset. ZSAD is challenging since the models need to generalize to anomalies across different domains where the appearance of foreground objects, abnormal regions, and background features can vary significantly.

Recently, large pre-trained vision-language models (VLMs) have demonstrated strong zero-shot recognition ability in various vision tasks. Particularly, being pre-trained using millions/billions of image-text pairs, CLIP [33] has been applied to empower various downstream tasks with its strong generalization capability. WinCLIP [17] is the first work to apply the CLIP model in the ZSAD line. It uses a large number of hand-crafted text prompts and extraction of window/patch/image-level features aligned with text to finish the ZSAD task. However, WinCLIP involves multiple forward passes of image patches for anomaly segmentation, which leads to inefficiency. Besides, WinCLIP only performs the testing of the CLIP model in the ZSAD without any fine-tuning using the anomaly detection task data, which also limits its performance in the ZSAD task. Subsequently, several works [8, 22, 32, 52] attempt to adapt the CLIP into the ZSAD task by fintuning. Specifically, AnomalyCLIP [52] learns object-agnostic text prompts that capture generic normality and abnormality in an image regardless of its foreground objects. ClipSAM [22] introduces

*Correponding Author

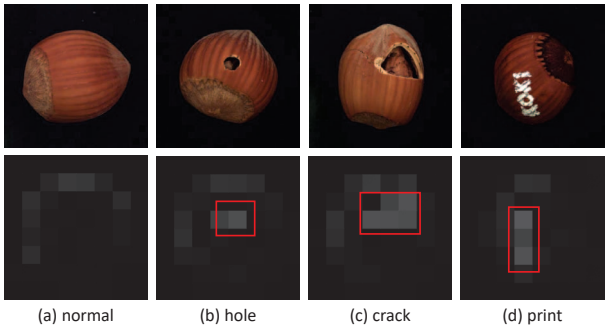

Figure 1. The local frequency statistics of images. The second row represents the local frequency statistics of non-overlapped image regions. There are four columns, standing for the normal object, object with a hole, object with a crack, and object with print, respectively. Please zoom in and pay attention to the red region (i.e. the anomaly region of the input image) for a better view.

a unified multiscale cross-modal interaction module for interacting language with visual features at multiple stages of the CLIP visual encoder to reason anomaly positions. VCP-CLIP [32] propose the Pre-VCP module and Post-VCP module to perform visual context prompting to activate CLIP's anomalous semantic perception ability. AdaCLIP [8] incorporates static and dynamic learnable prompts to adapt CLIP for ZSAD. However, all of the methods mentioned above overlook the utilization of frequency information of images.

In this paper, we find that frequency information could benefit the ZSAD task. Some properties of the anomaly area, such as appearance defects, can also be reflected based on its frequency information, since high-frequency components represent fine details (*e.g. soft borders*) that often exist in the anomaly area. For example, as shown in Figure 1, the frequency statistics in the red region are different from the frequency statistics in the surrounding regions of the input abnormal image and the corresponding region of the normal image. These specific frequency patterns could be helpful for anomaly detection.

To this end, we introduce a novel approach, namely the Frequency Enhanced CLIP (FE-CLIP), to inject frequency information into CLIP for accurate ZSAD across different domains. FE-CLIP takes advantage of two different but complementary frequency-aware clues, (1) Frequency-aware Feature Extraction (FFE) adapter, and (2) Local Frequency Statistics (LFS) adapter, in the visual encoder of CLIP, to deeply mine frequency information. Specifically, the FFE adapter uses DCT [1] to transform the image features into the frequency domain in order to mine the frequency information, then uses inverse DCT to transform back to the spatial domain. The FFE adapter describes the frequency-aware patterns in the spatial domain, but does not explicitly render the frequency information directly in the neural networks. Therefore, we use LFS to render the frequency information directly in the frequency domain. In each densely but regularly sampled local spatial patch, the statistics are gathered by counting the mean frequency responses. The averaged frequency statistics are a multi-channel spatial map, where the number of channels is identical to the number of channels in the spatial domain. The local frequency statistics also follow the spatial layouts as the features of the CLIP visual encoder, thus also enjoying effective representation learning powered by the CLIP visual encoder. Meanwhile, since the frequency-aware feature extraction and local frequency statistics are complementary to each other but both of them share inherently similar frequency-aware semantics, thus they can be progressively fused during the feature learning process.

Through comprehensive experiments, we show that the proposed FE-CLIP has good generalization across different domains and achieves superior zero-shot performance of detecting and segmenting anomalies in 10 datasets of highly diverse class semantics from various defect inspections and medical domains. Our contributions in this paper are summarized as follows:

1. We find that frequency information could benefit the ZSAD task, and propose a novel Frequency Enhanced CLIP (FE-CLIP) method to inject frequency information into CLIP for accurate ZSAD across different domains.
2. We propose two different but complementary frequency-aware adapters, Frequency-aware Feature Extraction (FFE) adapter, and Local Frequency Statistics (LFS) adapter, in the visual encoder of CLIP, to deeply mine the frequency patterns of anomaly region.
3. Extensive experiments show that the proposed FE-CLIP has good generalization across different domains and achieves superior zero-shot performance of detecting and segmenting anomalies in 10 datasets of highly diverse class semantics from various defect inspections and medical domains. Besides, the proposed FE-CLIP also achieves superior performance under the few-normal-shot anomaly detection settings.

2. Related Work

2.1. Vision-language modeling

Among recent successes of large pre-trained vision-language models (VLM) [2, 20, 33], CLIP [33] is the first to perform pre-training on web-scale image-text data, showing unprecedented generality: e.g., its language-driven zero-shot inference, improved both effective robustness [40] and perceptual alignment [12]. Many following VLM works explored large-scale pre-training in different aspects, e.g., scaling up data [20], efficient designs [21], multi-tasks [29], etc. To democratize large-scale VLM for the usages in different domains, a billion-scale data LAION5B [39], a

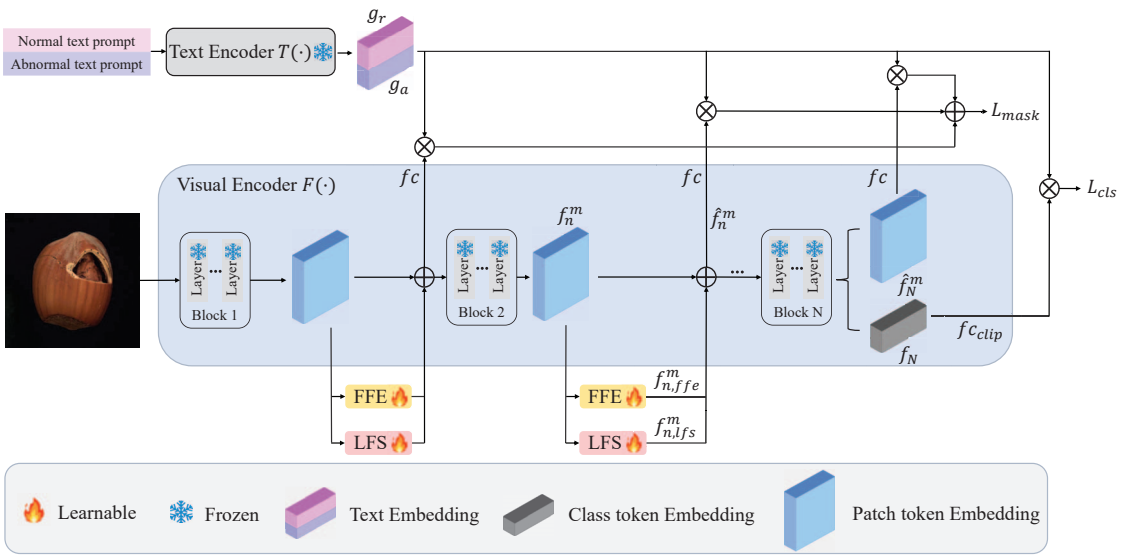

Figure 2. **The framework of the proposed FE-CLIP.** The text encoder and visual encoder are from the CLIP model, and the FFE and LFS are the proposed adapters. We only show the class token embedding after the last block of the CLIP visual encoder for convenience. Actually, the CLIP visual encoder can output the class token after every block.

code base of OpenCLIP with pre-trained models are open-sourced. Other works presented CLIP's promise in zero-/few-shot transfer to downstream tasks beyond classification [13, 36, 41, 46]. Good prompt engineering and tuning [33, 50] can non-trivially benefit generalization performances. Moreover, some other works [34, 48, 49] leverage the pre-trained CLIP for language-guided detection and segmentation with promising performances. In this paper, we mainly focus on adapting CLIP into the ZSAD task with superior performance by injecting the frequency information into the visual encoder of CLIP.

2.2. Zero-shot Anomaly detection

ZSAD relies on the model's strong transferability to handle unseen anomalies [3]. A very recent approach WinCLIP [17] presents a seminal work that leverages CLIP [33] for zero-shot anomaly classification and segmentation. WinCLIP proposes a compositional ensemble on state words and prompt templates and efficient extraction and aggregation of window/patch/image-level features aligned with text to finish the ZSAD task. To tackle this inefficiency, APRIL-GAN [9] introduces a fine-tuning strategy by adding extra learnable linear projection layers in multiple stages of the CLIP visual encoder to enhance the modeling of local visual semantics. AnomalyCLIP [52] learns object-agnostic text prompts that capture generic normality and abnormality in an image regardless of its foreground objects. ClipSAM [22] introduces a unified multiscale cross-modal interaction module for interacting language with visual features at multiple stages of the CLIP visual encoder to reason anomaly positions. Myriad [25] and AnomalyGPT [14] utilize large language models for zero-shot/few-shot anomaly detection. They employ an image decoder to provide fine-grained semantics and design a prompt learner to fine-tune the large language models using prompt embeddings. However, all of the methods mentioned above overlook the utilization of frequency information of images. In this paper, we find that frequency information could benefit the ZSAD task, and propose the FE-CLIP method which utilizes two different but complementary adapters, the FFE adapter and the LFS adapter, into the visual encoder of CLIP, to deeply mine the frequency information for ZSAD task.

3. Method

3.1. Preliminary

CLIP [33] is a large-scale pretraining method offering a joint vision-language representation. It consists of a text encoder $T(\cdot)$ and a visual encoder $F(\cdot)$. Both encoders are mainstream multi-block networks such as ViT [10]. Using text prompts is a typical way to acquire the embeddings of different classes for zero-shot classification. The text prompts usually combine a text prompt template Ω with the category name c, where the text prompt template Ω usually is defined as `A photo of a [cls]` and `[cls]` represents the target category name. The text prompts are passed through $T(\cdot)$ to obtain its corresponding textual embedding $g_c \in \mathbb{R}^D$ where D denotes the channel dimension. Then, an input image is passed through $F(\cdot)$ to get the visual representations, where the class token $f \in \mathbb{R}^D$ is the

global visual embedding of the input image, and patch tokens $f^m \in \mathbb{R}^{H\times W\times D}$ are treated as local visual embeddings of the input image. H and W denote the height and width of the patch tokens, respectively. CLIP can perform zero-shot recognition by measuring the similarity between textual and visual embeddings. Specifically, given a target class set $\mathcal{C}$ and an image, CLIP predicts the probability of the image belonging to the c category as follows:

$$p\left(y=c\right)=P\left(g_c,f\right)=\frac{\exp\left(<g_c,f>/\tau\right)}{\sum_{c\in\mathcal{C}}\exp\left(<g_c,f>\right)/\tau\right)} \quad (1)$$

where τ denotes the temperature and the operator $<\cdot,\cdot>$ represents the computation of cosine similarity. The computation can be extended from global visual embeddings to local visual embeddings to derive the corresponding segmentation maps $M_c \in \mathbb{R}^{H\times W}$, if each entry (i,j) is computed as $P\left(g_c, f^{m(i,j)}\right)$.

3.2. Frequency Enhanced CLIP

In this paper, we find that frequency information could benefit the ZSAD task, and propose a novel Frequency Enhanced CLIP (FE-CLIP) method to inject frequency information into CLIP [33] for accurate ZSAD across different domains. We inject the frequency information by the features of the CLIP visual encoder rather than the frequency of the input image, since the CLIP visual encoder is trained to take RGB image as input. Typically, the CLIP visual encoder comprises a series of block layers. From the bottom to the top of the layers, the visual encoder gradually learns the visual patterns at different levels of abstraction. As shown in Figure 2, we add Frequency-aware Feature Extraction (FFE) adapter and Local Frequency Statistics (LFS) adapter across multiple blocks levels in the visual encoder $F(\cdot)$ of CLIP, while keeping its original backbone unchanged, thus enabling frequency information injection at multiple feature levels. To be specific, assuming the visual encoder consists of N blocks ($N=4$), the features (i.e. the patch tokens) after the n-th block are denoted as $f_n^m \in \mathbb{R}^{H\times W\times D}$, and the class token after the n-th block is denoted as $f_n \in \mathbb{R}^D$, where $n=1,2,..,N$. Firstly, the features f_n^m are passed through an FFE adapter and an LFS adapter to get the features $f_{n,ffe}^m$ and $f_{n,lfs}^m$, respectively, which contain the frequency information. We can get the frequency-aware feature $\hat{f}_n^m$ as follows:

$$\hat{f}_n^m=\lambda(f_{n,ffe}^m+f_{n,lfs}^m)+(1-\lambda)f_n^m \quad (2)$$

with $\hat{f}_n^m$ serving as the input for the next block. The class token f_n will also be injected with frequency information after the next block due to the self-attention mechanism [44] in the blocks. We set $\lambda=0.1$ to preserve the original knowledge of the CLIP visual encoder. Then, the normal text prompt and abnormal text prompt are passed through text encoder $T(\cdot)$ to get corresponding textual embedding g_r and g_a, respectively. g_r and g_a constitute the target class set $\mathcal{C}$. We simply use `A photo of a normal object` and `A photo of a damaged object` for the normal text prompt and abnormal text prompt, respectively, rather than the complex compositional ensemble on state words and prompt templates used in WinCLIP [17], since we find that the simple text prompt also works well. Finally, we compute the abnormal segmentation map $M_{a,n} \in \mathbb{R}^{H\times W}$ based on the equation 1, where each entry (i,j) is computed as $P\left(g_a, fc(\hat{f}_n^{m(i,j)})\right)$. The fc means that we use a single learnable fc to align the dimension of visual features $\hat{f}_n^m$ to the dimension of text features. The abnormal score $S_{a,n}$ is also computed based on the equation 1 as $P\left(g_a, fc_{clip}(f_n)\right)$. The fc_{clip} denotes the frozen project layer of the visual encoder in CLIP to align the dimension of class token f_n and the dimension of text features. The final prediction of the abnormal score S_a and abnormal segmentation map M_a are computed by averaging $S_{a,n}$ and $M_{a,n}$ across N blocks, respectively, as follows:

$$S_a=\frac{1}{N}\sum_{n=1}^{N}S_{a,n}, \quad M_a=\frac{1}{N}\sum_{n=1}^{N}M_{a,n} \quad (3)$$

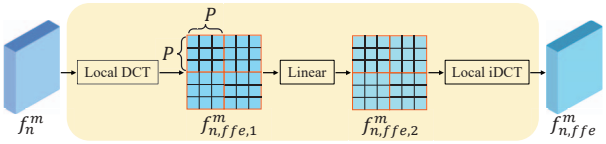

Figure 3. **The proposed Frequency-aware Feature Extraction (FFE) adapter.**

3.3. Frequency-aware Feature Extraction Adapter

The specific frequency patterns could be beneficial for anomaly detection. To mine the frequency information, we propose the Frequency-aware Feature Extraction (FFE) adapter in the visual encoder of CLIP. Specifically, given the patch tokens of the n-th block $f_n^m \in \mathbb{R}^{H\times W\times D}$, we first split the f_n^m into non-overlapped windows with total $(H/P)\times(W/P)$ windows and each window containing $P\times P$ patch tokens (P is set to 3 by default). As shown in Figure 3, we apply DCT [1] transformation to each window to extract the DCT features $f_{n,ffe,1}^m$ which represents the frequency information. Then, we pass $f_{n,ffe,1}^m$ through one linear layer followed by one $GELU(\cdot)$ activation layer to further extract the frequency features $f_{n,ffe,2}^m$. Finally, we apply the inverse DCT transformation to $f_{n,ffe,2}^m$ to get the frequency-aware features $f_{n,ffe}^m$ in the spatial domain. The FFE adapter describes the frequency-aware patterns in the spatial domain, but does not explicitly render the frequency information directly in the neural networks. In the next sub-section, we propose an adapter that renders the fre-

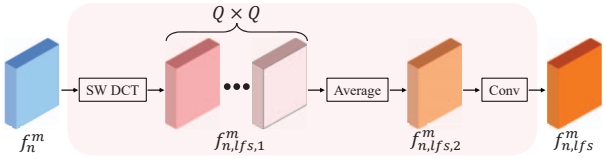

Figure 4. **The proposed Local Frequency Statistics (LFS) adapter.**

quency information directly in the neural networks.

3.4. Local Frequency Statistics Adapter

The aforementioned FFE adapter has provided frequency-aware representation, but it has to represent the frequency-aware patterns back into the spatial domain, thus failing to directly utilize the frequency patterns in the frequency domain. Therefore, we propose the Local Frequency Statistics (LFS) adapter in the visual encoder of CLIP to render the frequency patterns directly in the frequency domain. As shown in Figure 4, given the patch tokens of the n-th block $f_n^m \in \mathbb{R}^{H\times W\times D}$, we first apply a Sliding Window DCT (SW DCT) on f_n^m (i.e., taking DCT [1] densely on sliding windows of the features) to extract the localized frequency responses $f_{n,lfs,1}^m$. The size of sliding windows is $Q\times Q$ (Q is set to 3 by default). Therefore, there are $Q\times Q$ group frequency responses in $f_{n,lfs,1}^m$, where the first one represents the low-frequency responses and the last one represents the high-frequency responses. Then we count the mean of $f_{n,lfs,1}^m$ across $Q\times Q$ groups to get the mean frequency responses $f_{n,lfs,2}^m$. The $f_{n,lfs,2}^m$ is a multi-channel spatial map that shares the same layout as f_n^m. The LFS adapter outputs $f_{n,lfs}^m$ after $f_{n,lfs,2}^m$ being passed through a convolutional layer followed by one $GELU(\cdot)$ activation layer. The LFS adapter provides a localized aperture to detect detailed abnormal frequency distributions.

3.5. Loss Functions

Let S_{gt} denote the ground-truth image label where 1 represents the abnormal image and 0 represents the normal image. The loss for the anomaly classification is defined as follows:

$$L_{cls} = -\frac{1}{N}\sum_{N}^{n=1}[S_{gt}log(S_{a,n}) + (1-S_{gt})log(1-S_{a,n})] \tag{4}$$

where log denotes the log operation. Similarly, let M_{gt} denote the groud-truth pixel label of the input image, where 1 means abnormal pixel region and 0 means normal pixel region. The loss for the anomaly segmentation is defined as follows:

$$L_{mask} = \frac{1}{N}\sum_{N}^{n=1}[Focal(Up(M_{a,n}), M_{gt}) + Dice(Up(M_{a,n}), M_{gt})] \tag{5}$$

where $Focal(\cdot,\cdot)$ and $Dice(\cdot,\cdot)$ denote the focal loss [26] and dice loss [23], respectively. The operator $Up(\cdot)$ represents the unsampling operation. During training, both the visual encoder and the text encoder of CLIP [33] are Frozen. Only the FFE adapters and LFS adapters are optimized by the loss function $L_{total} = L_{cls} + L_{mask}$.

4. Experiments

4.1. Datasets and Evaluation Metrics

To verify the effectiveness of the proposed method FE-CLIP, we conduct comprehensive experiments across 10 real-world anomaly detection datasets, covering various industrial inspection scenarios and medical imaging domains. In industrial inspection, we use MVTec AD [4], VisA [54], MPDD [18], BTAD [30], DAGM [45], and DTD-Synthetic [3]. In medical imaging, we use colon polyp detection datasets CVC-ClinicDB [6] and Kvasir [19], brain tumor detection datasets BrainMRI [38] and Br35H [15].

As for the evaluation metrics, following previous works [17, 52], we use Area Under the Receiver Operating Characteristic curve (AUROC). Additionally, Average Precision (AP) for Zero-Shot Anomaly Detection (ZSAD) and PRO [5] for Zeor-Shot Anomaly Segmentation (ZSAS) are also used to provide a more in-depth analysis of the results.

4.2. Implementation Details

We use the publicly available CLIP model (VIT-L/14@336px) as our backbone. Model parameters of CLIP are all frozen, and only the parameters of FFE adapters and LFS adapters are learnable. Following AnomalyCLIP [52], we fine-tune FE-CLIP using the test data of MVTec AD and evaluate the ZSAD performance on other datasets. As for MVTec AD, we fine-tune FE-CLIP on the test data of VisA. The proposed FE-CLIP is trained by 9 epochs with Adam optimizer. The learning rate is set to 5e-4 and the total batch size is 16. All experiments are conducted in PyTorch-1.13.0 with four NVIDIA RTX 3090 24GB GPUs. All experiment results are averaged under five runs. We report dataset-level results, which are averaged across their respective sub-datasets.

4.3. Main Results

ZSAD performance on diverse anomaly scenarios. Table 1 shows the ZSAD results of FE-CLIP with other competing methods over six industrial defect datasets and two medical domain datasets. As shown in Table 1, FE-CLIP can achieve higher AUROC and AP for the ZSAD task on most industrial inspection datasets and medical datasets, such as MVTec AD, DTD-Synthetic, and BrainMRI, when compared with other competing methods, such as WinCLIP [17], AnomalyCLIP [52] and AdaCLIP [8]. This shows that (1) the frequency information can benefit the ZSAD task

Methods	Industrial Defects						Medical Anomalies	
	MVTec AD	VisA	MPDD	BTAD	DAGM	DTD-Synthetic	BrainMRI	Br35H
CLIP [33]	74.1 / 87.6	66.4 / 71.5	54.3 / 65.4	34.5 / 52.5	79.6 / 59.0	71.6 / 85.7	73.9 / 81.7	78.4 / 78.8
CoOp [50]	88.8 / 94.8	62.8 / 68.1	55.1 / 64.2	66.8 / 77.4	87.5 / 74.6	- / -	61.3 / 44.9	86.0 / 87.5
WinCLIP [17]	91.8 / 96.5	78.1 / 81.2	63.6 / 69.9	68.2 / 70.9	91.8 / 79.5	93.2 / 92.6	86.6 / 91.5	80.5 / 82.2
APRIL-GAN [9]	86.1 / 93.5	78.0 / 81.4	78.0 / 81.4	73.6 / 68.6	94.4 / 83.8	86.4 / 95.0	89.3 / 90.9	93.1 / 92.9
AnomalyCLIP [52]	91.5 / 96.2	82.1 / 85.4	77.0 / 82.0	88.3 / 87.3	97.5 / 92.3	93.5 / 97.0	90.3 / 92.2	94.6 / 94.7
AdaCLIP [8]	89.2 / -	85.8 / -	76.0 / -	88.6 / -	99.1 / -	95.5 / -	94.8 / -	97.7 / -
FE-CLIP	91.9 / 96.5	84.6 / 86.6	78.0 / 82.6	90.3 / 90.0	97.5 / 92.3	98.3 / 99.4	94.8 / 93.8	96.8 / 93.8

Table 1. AUROC / AP results for Zero-Shot Anomaly Detection (ZSAD) task on eight real-world anomaly detection datasets. Best results are highlighted in red.

Methods	Industrial Defects						Medical Anomalies	
	MVTec AD	VisA	MPDD	BTAD	DAGM	DTD-Synthetic	CVC-ClinicDB	Kvasir
CLIP [33]	38.4 / 11.3	46.6 / 14.8	62.1, 33.0	30.6 / 4.4	30.6 / 4.4	33.9 / 12.5	47.5 / 18.9	44.6 / 17.7
CoOp [50]	33.3 / 6.7	24.2 / 3.8	15.4 / 2.3	15.4 / 2.3	17.5 / 2.1	- / -	17.5 / 2.1	44.1 / 3.5
WinCLIP [17]	85.1 / 64.6	79.6 / 56.8	79.6 / 56.8	72.7 / 27.3	87.6 / 65.7	83.9 / 57.8	51.2 / 13.8	69.7 / 24.5
APRIL-GAN [9]	87.6 / 84.0	94.2 / 86.8	94.1 / 83.2	60.8 / 25.0	82.4 / 66.2	95.3 / 86.9	80.5 / 60.7	75.0 / 36.2
AnomalyCLIP [52]	91.1 / 81.4	95.5 / 87.0	96.5 / 88.7	94.2 / 74.8	95.6 / 91.0	97.9 / 92.3	82.9 / 67.8	78.9 / 45.6
AdaCLIP [8]	88.7 / -	95.5 / -	96.1 / -	92.1 / -	91.5 / -	97.9 / -	84.4 / -	- / -
VCP-CLIP [32]	92.0 / 87.3	95.7 / 90.7	- / -	94.1 / 74.6	99.4 / 98.3	- / -	- / -	- / -
FE-CLIP	92.6 / 88.3	95.9 / 92.8	97.0 / 90.5	95.6 / 80.4	98.5 / 96.6	99.0 / 97.4	84.5 / 69.2	79.8 / 60.6

Table 2. AUROC / PRO results for Zero-Shot Anomaly Segmentation (ZSAS) task on eight real-world anomaly detection datasets. Best results are highlighted in red. Note that the image-level medical AD datasets do not contain segmentation ground truth, so the pixel-level medical AD datasets are different from the image-level datasets.

due to the superior performance of FE-CLIP, and (2) the frequency information helps FE-CLIP achieve good generalization across datasets from different domains.

ZSAS performance on diverse anomaly scenarios. We show the ZSAS results of FE-CLIP with other competing methods over six industrial defect datasets and two medical domain datasets in Table 2. Table 2 shows that FE-CLIP achieves higher AUROC and PRO for the ZSAS task on almost all of the industrial inspection datasets and medical datasets, except for the DAGM dataset, when compared with other competing methods, such as WinCLIP [17], AnomalyCLIP [52], AdaCLIP [8] and VCP-CLIP [32]. This further demonstrates that (1) the frequency information can benefit the ZSAS task and (2) the frequency information helps FE-CLIP achieve good generalization across datasets from different domains.

Discussion. We also note that the FE-CLIP can defeat almost all of the competing methods on the ZSAS task, while FE-CLIP only performs better than other methods over most of the datasets on ZSAD task. The reason may be that the frequency information can be directly injected into the patch tokens (i.e. the feature map used for the ZSAS task) through the FFE adapter and LFS adapter, while the frequency information can only be implicitly involved in the global class token (i.e. the feature vector used for the ZSAD task) from the patch tokens based on the self-attention mechanism.

Overall, the proposed FE-CLIP achieves superior zero-shot performance of detecting and segmenting anomalies in datasets of highly diverse class semantics from various defect inspection and medical imaging domains.

4.4. Ablation Study

The effectiveness of FFE and LFS We conduct ablation experiments on MVTec AD dataset and VisA dataset to validate the effectiveness of the proposed FFE adapter and LFS adapter. The results are shown in Table 3. The "CLIP [33] + Conv Adapter" means we add N conv adapters, where each of them consists of one convolutional layer followed by one $GELU(\cdot)$ activation layer to keep the same learnable parameters with the LFS adapter into the visual encoder of CLIP. It can be concluded that both the FFE adapter and LFS adapter can achieve better performance than the conv adapter in MVTec AD dataset and VisA dataset across the ZSAD task and ZSAS task. This also shows that the frequency information can benefit the ZSAD task and ZSAS task. When combined the FFE adapter with the LFS adapter, the FE-CLIP achieves higher performance than the FFE adapter and LFS adapter. This demonstrates that the FFE adapter and LFS adapter can mine different but complementary frequency information to further boost the performance of model on the ZSAD and ZSAS tasks.

Methods	MVTec AD		VisA	
	Pixel-level	Image-level	Pixel-level	Image-level
CLIP[33] + Conv Adapter	89.0 / 85.1	87.8 / 94.5	94.8 / 89.8	82.3 / 83.5
CLIP[33] + FFE Adapter	90.6 / 86.2	90.0 / 95.4	95.1 / 91.3	83.3 / 85.0
CLIP[33] + LFS Adapter	91.5 / 87.1	90.6 / 95.7	95.5 / 91.8	83.9 / 85.7
FE-CLIP	92.6 / 88.3	91.9 / 96.5	95.9 / 92.8	84.6 / 86.6

Table 3. Module ablation on the MVTec AD and the VisA datasets. The metric of Pixel-level is AUROC / PRO for ZSAS task, and the metric of Image-level is AUROC / AP for ZSAD task. Best results are highlighted in red.

Methods	MVTec AD		VisA	
	Pixel-level	Image-level	Pixel-level	Image-level
FFT	92.8 / 88.3	92.0 / 96.2	95.5 / 93.0	84.1 / 86.8
DCT	92.6 / 88.3	91.9 / 96.5	95.9 / 92.8	84.6 / 86.6

Table 4. Compared with using FFT to extract the frequency information in the FFE and LFS modules. The metric of Pixel-level is AUROC / PRO for ZSAS task, and the metric of Image-level is AUROC / AP for ZSAD task. Best results are highlighted in red.

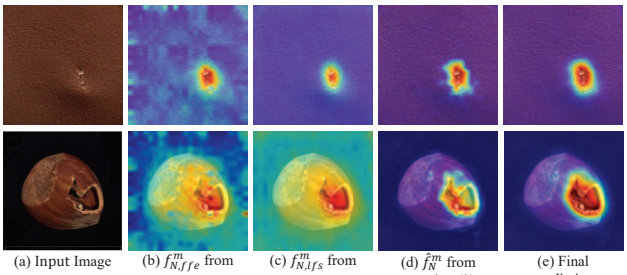

Figure 5. The visualization of the learned frequency features. Please zoom in for a better view.

Would FFT work instead of DCT We attempt to use FFT to replace the DCT in the FFE and LFS modules to extract the frequency information. However, there are real values and imaginary values after FFT. To deal with this, the real values and imaginary values are separately processed and then combined to transform back to the spatial domain in the FFE module. As for the LFS module, the averaged local frequency statistics are computed from the real values and imaginary values, respectively, to get the averaged real local frequency statistics and the averaged imaginary local frequency statistics, and then they are summed to get the final averaged local frequency statistics. The results are shown in Table 4. This shows that using FFT achieves similar results with DCT, which means that the key is to inject the frequency information rather than the way of frequency transformation. We choose DCT since DCT is more convenient to extract frequency.

4.5. Analysis

Time cost. We report the inference time of APRIL-GAN, AnomalyCLIP, VCP-CLIP, AdaCLIP, and FE-CLIP with PyTorch-1.13.0 on one NVIDIA RTX 3090 24GB GPU. The inference time is 105, 120, 128, 160, and 125 ms, respectively. FE-CLIP achieves a better trade-off between performance and speed.

Visualization of learned frequency features. To show why frequency information is effective for the ZSAD task, we visualize the frequency-aware features from the last block of the visual encoder of CLIP as example, where the FFE and LFS adapters are added. As shown in Figure 5, both the features $f^m_{N,ffe}$ from FFE adapter and the features $f^m_{N,lfs}$ from LFS adapter highlight the abnormal region of images. By injecting the features $f^m_{N,ffe}$ and $f^m_{N,lfs}$ into the visual encoder of CLIP, the obtained features $\hat{f}^m_N$ can successfully localize the main abnormal region, in favor of the final prediction of the model.

When frequency information can be beneficial. Some properties of anomaly area, such as appearance defects, can be reflected based on its frequency information, since high-frequency components represent fine details (e.g. soft borders) that often exist in anomaly area. In these cases, These specific frequency patterns could be beneficial for anomaly detection. For example, as shown in Figure 1, the frequency statistics in the red region are different from the frequency statistics in the surrounding regions of the input abnormal image and the corresponding region of the normal image.

Let's consider images of normal dumplings and images of squashed dumplings, where squashed dumplings are anomalies. In this case, if the squashed dumplings reveal the meat filling, the frequency of squashed dumplings could still stand out and frequency information could be helpful to detect anomalies. However, if the dumplings are squashed to be more like a meat pie without revealing the meat filling, the frequency of this squashed dumpling might not stand out and frequency information might not work well.

We validate the FE-CLIP across different domains with 10 datasets of highly diverse class semantics from various industrial inspections and medical domains, and the results of FE-CLIP on the datasets of medical domains are trained

Setup	Methods	MVTec AD		VisA	
		Pixel-level	Image-level	Pixel-level	Image-level
1-shot	PatchCore [37]	92.0 / 79.7	83.4 / 92.2	95.4 / 80.5	79.9 / 82.8
	WinCLIP [17]	95.2 / 87.1	93.1 / 96.5	96.4 / 85.1	83.8 / 85.1
	AnomalyGPT [14]	95.3 / -	94.1 / -	96.2 / -	87.4 / -
	PromptAD [24]	95.9 / -	94.6 / -	96.7 / -	86.9 / -
	FE-CLIP	96.1 / 90.9	95.6 / 98.1	97.9 / 93.9	88.7 / 90.5
2-shot	PatchCore [37]	93.3 / 82.3	86.3 / 93.8	96.1 / 82.6	81.6 / 84.8
	WinCLIP [17]	96.0 / 88.4	94.4 / 97.0	96.8 / 86.2	84.6 / 85.8
	AnomalyGPT [14]	95.6 / -	95.5 / -	96.4 / -	88.6 / -
	PromptAD [24]	96.2 / -	95.7 / -	97.1 / -	88.3 / -
	InCTRL [53]	- / -	94.0 / 96.9	- / -	85.8 / 87.7
	FE-CLIP	96.5 / 91.5	96.2 / 98.3	98.2 / 94.3	90.3 / 91.6
4-shot	PatchCore [37]	94.3 / 84.3	88.8 / 94.5	96.8 / 84.9	85.3 / 87.5
	WinCLIP [17]	96.2 / 89.0	95.2 / 97.3	97.2 / 87.6	87.3 / 88.8
	AnomalyGPT [14]	96.2 / -	96.3 / -	96.7 / -	90.6 / -
	PromptAD [24]	96.5 / -	96.6 / -	97.4 / -	89.1 / -
	InCTRL [53]	- / -	94.5 / 97.2	- / -	87.7 / 90.2
	FE-CLIP	96.7 / 91.7	96.6 / 98.4	98.4 / 95.0	90.7 / 91.9

Table 5. Compared with other competing methods under few-normal-shot anomaly detection setting. The metric of Pixel-level is AUROC / PRO for the few-normal-shot anomaly segmentation task, and the metric of Image-level is AUROC / AP for the few-normal-shot anomaly detection task. Best results are highlighted in red.

by the industrial inspection MVTec AD dataset, as shown in Table 1 and Table 2. The superior performance of FE-CLIP demonstrates that frequency information could be beneficial for many cases of anomaly detection, and the proposed FE-CLIP has good generalization across different domains.

4.6. FE-CLIP with Few Normal Shots

The performance of FE-CLIP can also be further boosted when equipped with few-shot normal images, just like WinCLIP [17] and AnomalyGPT [14]. We adopt the same way as AnomalyGPT to compute the anomaly segmentation map with few-shot normal images. In short words, we use the same visual encoder of FE-CLIP to extract intermediate patch-level features from normal images and store them in memory banks, then calculate the cosine similarity distance between each patch token of input image and its most similar counterpart in the memory banks to obtain the anomaly segmentation map S_{a,few_shot}. Finally, we add the anomaly segmentation map S_a from the FE-CLIP model with S_{a,few_shot} to get the final anomaly segmentation map, and add the anomaly score M_a from the FE-CLIP model with the max value of S_{a,few_shot} to get the final anomaly score. Please refer to [14] for more technical details.

We compare the FE-CLIP with other competing methods, such as AnomalyGPT [14], PromptAD [24], and InCTRL [53], under the few-shot anomaly detection setting. As shown in Table 5, the FE-CLIP outperforms all other competing methods across different metrics, including AUROC and PRO metrics for the anomaly detection task, and AUROC and AP metrics for the anomaly segmentation task, on both datasets. The experiments under few-normal-shot setting further demonstrate the superiority of the FE-CLIP.

5. Conclusion

In this paper, we find that frequency information can benefit the ZSAD task. To this end, we propose FE-CLIP, which adds the proposed FFE and LFS adapters into the visual encoder of CLIP, to deeply mine the frequency patterns for the visual features. Extensive experiments on 10 datasets of highly diverse class semantics from various defect inspections and medical domains demonstrate that (1) the frequency information can benefit the ZSAD task, and (2) the frequency information can help the model have good generalization across datasets from different domains. We also conduct experiments under the few-normal-shot anomaly detection setting to further validate the superiority of the proposed FE-CLIP. We hope the community of anomaly detection will be aware of the effectiveness of frequency information from our work and explore more ways to utilize the frequency information for anomaly detection in the future.

Acknowledgements

This work was supported by the Anhui Provincial Science and Technology Major Project (No. 2023z020006).

References

[1] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1): 90–93, 1974. 2, 4, 5

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2

[3] Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5564–5572, 2023. 3, 5

[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. 1, 5

[5] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020. 1, 5

[6] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015. 5

[7] Tri Cao, Jiawen Zhu, and Guansong Pang. Anomaly detection under distribution shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6511–6523, 2023. 1

[8] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, 2024. 1, 2, 5, 6

[9] 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. 3, 6

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[11] Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37, 2021. 1

[12] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021. 2

[13] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 3

[14] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1932–1940, 2024. 3, 8

[15] A. Hamada. Br35h: Brain tumor detection 2020. In *https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection*, 2020. 5

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

[17] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023. 1, 3, 4, 5, 6, 8

[18] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*, pages 66–71. IEEE, 2021. 5

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

[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. 2

[21] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[22] Shengze Li, Jianjian Cao, Peng Ye, Yuhan Ding, Chongjun Tu, and Tao Chen. Clipsam: Clip and sam collaboration for zero-shot anomaly segmentation. *arXiv preprint arXiv:2401.12665*, 2024. 1, 3

[23] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*, 2019. 5

[24] Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Promptad: Learning prompts with only normal samples for few-shot anomaly

detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 8

[25] Yuanze Li, Haolin Wang, Shihao Yuan, Ming Liu, Debin Zhao, Yiwen Guo, Chen Xu, Guangming Shi, and Wangmeng Zuo. Myriad: Large multimodal model by applying vision experts for industrial anomaly detection. *arXiv preprint arXiv:2310.19070*, 2023. 3

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[27] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023. 1

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

[29] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[30] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE, 2021. 5

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

[32] Zhen Qu, Xian Tao, Mukesh Prasad, Fei Shen, Zhengtao Zhang, Xinyi Gong, and Guiguang Ding. Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. In *European Conference on Computer Vision*, 2024. 1, 2, 6

[33] 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, 2021. 1, 2, 3, 4, 5, 6, 7

[34] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091, 2022. 3

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

[36] 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. 3

[37] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022. 1, 8

[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. 5

[39] 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. 2

[40] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020. 2

[41] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928, 2022. 3

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

[43] Yu Tian, Fengbei Liu, Guansong Pang, Yuanhong Chen, Yuyuan Liu, Johan W Verjans, Rajvinder Singh, and Gustavo Carneiro. Self-supervised pseudo multi-class pre-training for unsupervised anomaly detection and segmentation in medical images. *Medical image analysis*, 90:102930, 2023. 1

[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. 4

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

[46] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 3

[47] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class

anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022. 1

[48] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 3

[49] 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. 3

[50] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3, 6

[51] Qihang Zhou, Shibo He, Haoyu Liu, Tao Chen, and Jiming Chen. Pull & push: Leveraging differential knowledge distillation for efficient unsupervised anomaly detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 1

[52] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *The thirteenth International Conference on Learning Representations*, 2024. 1, 3, 5, 6

[53] Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 8

[54] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022. 5