Adapting Visual-Language Models for Generalizable Anomaly Detection in Medical Images

Chaoqin Huang[1,2,3]*, Aofan Jiang[1,3]*, Jinghao Feng[1,3], Ya Zhang[1,3], Xinchao Wang[2,†], Yanfeng Wang[1,3,†]
[1] Shanghai Jiao Tong University, [2] National University of Singapore
[3] Shanghai Artificial Intelligence Laboratory
{huangchaoqin,stillunnamed,fjh1345528968,ya_zhang,wangyanfeng622}@sjtu.edu.cn; {xinchao}@nus.edu.sg

Abstract

Recent advancements in large-scale visual-language pre-trained models have led to significant progress in zero-/few-shot anomaly detection within natural image domains. However, the substantial domain divergence between natural and medical images limits the effectiveness of these methodologies in medical anomaly detection. This paper introduces a novel lightweight multi-level adaptation and comparison framework to repurpose the CLIP model for medical anomaly detection. Our approach integrates multiple residual adapters into the pre-trained visual encoder, enabling a stepwise enhancement of visual features across different levels. This multi-level adaptation is guided by multi-level, pixel-wise visual-language feature alignment loss functions, which recalibrate the model's focus from object semantics in natural imagery to anomaly identification in medical images. The adapted features exhibit improved generalization across various medical data types, even in zero-shot scenarios where the model encounters unseen medical modalities and anatomical regions during training. Our experiments on medical anomaly detection benchmarks demonstrate that our method significantly surpasses current state-of-the-art models, with an average AUC improvement of 6.24% and 7.33% for anomaly classification, 2.03% and 2.37% for anomaly segmentation, under the zero-shot and few-shot settings, respectively. Source code is available at: https://github.com/MediaBrain-SJTU/MVFA-AD

1. Introduction

Medical anomaly detection (AD), which focuses on identifying unusual patterns in medical data, is central to preventing misdiagnoses and facilitating early interventions [14, 47, 61]. The vast variability in medical images, both in terms of modalities and anatomical regions, necessitates a model that is versatile across various data types. The few-shot AD approaches [12, 22, 46, 57] strive to attain model generalization with scarce training data, embodying a preliminary attempt for a universal AD model, despite the need for lightweight re-training [12, 46, 57] or distribution adjustment [22] for each new AD task.

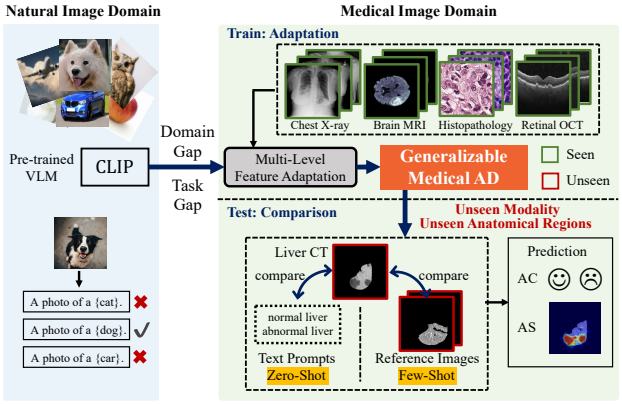


Figure 1. The overview of adaptation in pre-trained visual-language models for zero-/few-shot medical anomaly classification (AC) and anomaly segmentation (AS).

Contemporary large-scale pre-trained visual-language models (VLMs) have recently paved the way for robust and generalizable anomaly detection. A notable initial effort is to directly adopt CLIP [38], a representative open-source VLM for natural imagery, for AD, simply by carefully crafting artificial text prompts [26]. By further employing annotated training data, Chen *et al.* [9] introduces extra linear layers to map the image features to the joint embedding space to the text features, facilitating their comparison. Despite the promise of the above two approaches, their extension to the medical domain has not been explored.

This paper attempts to develop a universal generalizable AD model for medical images, designed to be adaptable to previously unseen modalities and anatomical regions. The

*Equal Contribution
†Corresponding authors: Yanfeng Wang and Xinchao Wang

creation of such a model holds significant practical importance, but tailoring the CLIP model for this purpose presents a triad of challenges. Firstly, re-purposing CLIP for AD signifies a substantial shift in task requirements. The visual encoder in CLIP is known to primarily capture image semantics, yet a universal AD model must discern irregularities across diverse semantic contexts. Secondly, the transition from using CLIP in the realm of natural imagery to medical imagery constitutes a significant domain shift. Finally, the task of extending the AD model's applicability to unencountered imaging modalities and anatomical regions during the training phase is notably demanding.

This paper proposes a lightweight *Multi-level Adaptation and Comparison* framework to re-purpose CLIP for AD in medical images as shown in Figure 1. A multi-level visual feature *adaptation* architecture is designed to align CLIP's features to the requirements of AD in medical contexts. The process of visual feature adaptation merges adapter tuning with multi-level considerations. This is achieved by integrating multiple residual adapters into the pre-trained visual encoder. This stepwise enhancement of visual features across different levels is guided by multi-level, pixel-wise visual-language feature alignment loss functions. These adapters recalibrate the model's focus, shifting it from object semantics to identifying anomalies in images, utilizing text prompts that broadly categorize images as 'normal' or 'anomalous'. During testing, *comparison* is performed between the adapted visual features and text prompt features, and additional referenced image features if available, enabling the generation of multi-level anomaly score maps.

The methods are evaluated on a challenging medical AD benchmark, encompassing datasets from five distinct medical modalities and anatomical regions: brain MRI [1, 2, 35], liver CT [7, 29], retinal OCT [21, 28], chest X-ray [51], and digital histopathology [4]. Our method outperforms state-of-the-art approaches, showcasing an average improvement of 6.24% and 7.33% in anomaly classification, and 2.03% and 2.37% in anomaly segmentation under zero-shot and few-shot scenarios, respectively.

The main contributions are summarized below:

- A novel multi-level feature adaptation framework is proposed, which is, to the best of our knowledge, the first attempt to adapt pre-trained visual-language models for medical AD in zero-/few-shot scenarios.
- Extensive experiments on a challenging benchmark for AD in medical images have demonstrated its exceptional generalizability across diverse data modalities and anatomical regions.

2. Related Works

Vanilla Anomaly Detection. Given the limited availability and high cost of abnormal images, a portion of current research on AD focuses on unsupervised methods relying exclusively on normal images [6, 10, 17, 19, 31, 41, 43, 44, 50, 54, 58]. Approaches such as PatchCore [41] create a memory bank of normal embeddings and detect anomalies based on the distance from a test sample to the nearest normal embedding. Another method, CflowAD [19], projects normal samples onto a Gaussian distribution using normalizing flows. However, relying solely on normal samples can result in an ambiguous decision boundary and reduced discriminability [5]. In practical scenarios, a small number of anomaly samples are usually available, and these can be used to enhance detection effectiveness.

Zero-/Few-Shot Anomaly Detection. The utilization of a few known anomalies during training can present challenges, potentially biasing the model and hindering generalization to unseen anomalies. DRA [12] and BGAD [57] introduce methods to mitigate this issue. Beyond simply maximizing the separation of abnormal features from normal patterns [37, 42], DRA [12] learns disentangled representations of anomalies to enable generalizable detection, accounting for unseen anomalies. BGAD [57] proposes a boundary-guided semi-push-pull contrastive learning mechanism to further alleviate the bias issue. Recent advancements like WinCLIP [26] explore the use of foundation models for zero-/few-shot AD, leveraging language to assist in AD. Building upon WinCLIP, April-GAN [9] maps visual features extracted from CLIP to the linear space where the text features are located, supervised by pixel-level annotated data. This paper concentrates on medical AD, a more challenging area than traditional industrial AD due to the larger gap between different data modalities.

Medical Anomaly Detection. Current medical AD methods typically treat AD as a one-class classification issue, relying on normal images for training [3, 8, 27, 61, 64–66]. These methods, which identify anomalies as deviations from the normal distribution, often require a large number of normal samples per class, making them impractical for real-world diagnosis. Many of these techniques are designed for a particular anatomical region [13, 55] or are restricted to handling one specific data type per model [24, 30, 53]. These methods often fall short in terms of generalizing across diverse data modalities and anatomical regions [62], a pivotal aspect our paper aims to address.

Visual-Language Modeling. Recently, VLMs have witnessed substantial advancements, applied to many different scenarios [20, 32, 34, 56, 59, 60]. Trained on a vast amount of image-text data, CLIP [38] excels in generalizability and robustness, notably enabling language-driven zero-shot inference [16, 48]. To broaden the application of VLMs, resources like the extensive LAION-5B dataset [45] and the OpenCLIP codebase [25] have been made available openly. Subsequent research has underscored CLIP's potential for zero-/few-shot transfer to downstream tasks beyond mere classification [18, 40, 67]. Several studies [23, 39, 63] have

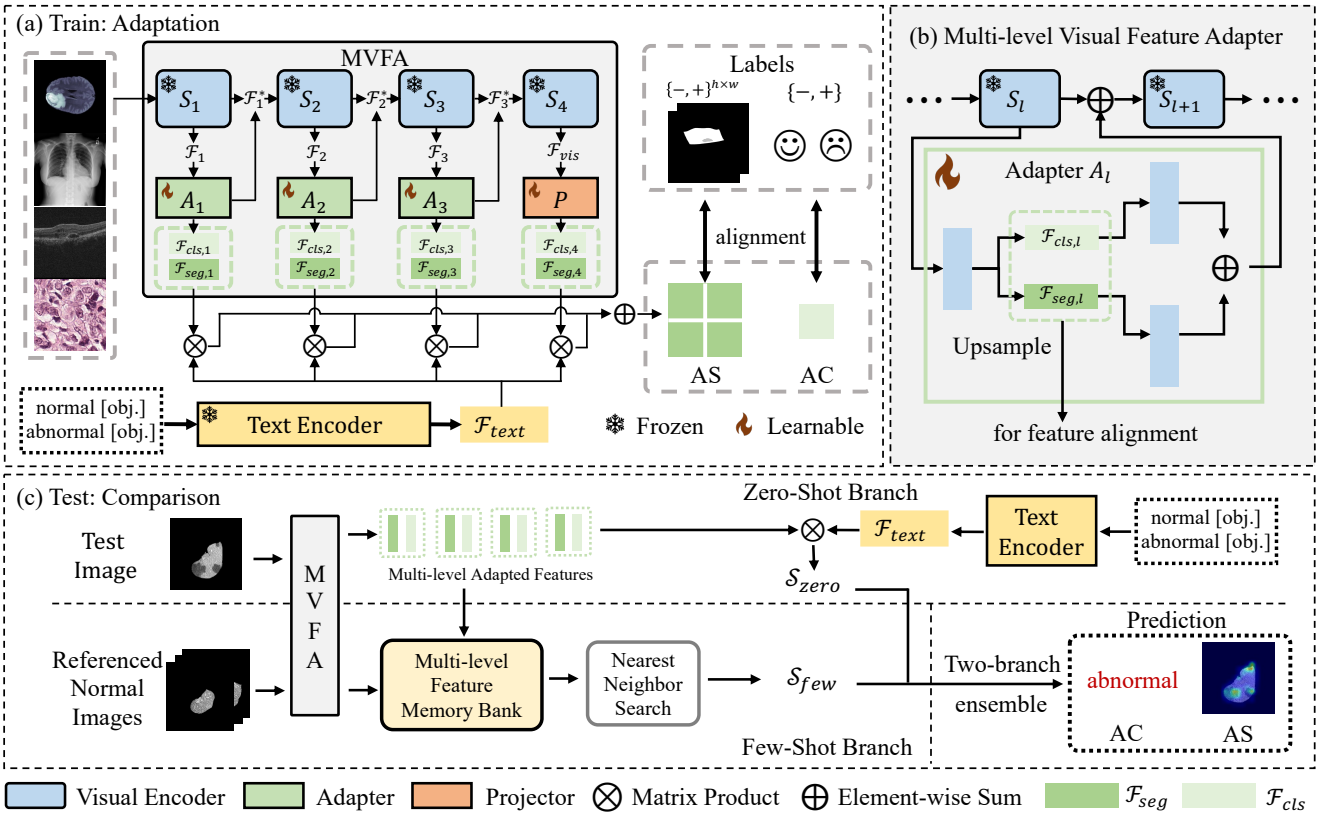


Figure 2. The architecture of multi-level adaptation and comparison framework for zero-/few-shot medical anomaly detection.

leveraged pre-trained CLIP models for language-guided detection and segmentation, achieving promising outcomes. This research extends the application of VLMs, initially trained on natural images, to AD in medical images, introducing a unique approach of multi-level visual feature adaptation and comparison framework.

3. Problem Formulation

We aim to adapt a visual-language model, initially trained on natural images (denoted as $\mathcal{M}_{nat}$), for anomaly detection (AD) in medical images, resulting in a medically adapted model ($\mathcal{M}_{med}$). This adaptation utilizes a medical training dataset $\mathcal{D}_{med}$, which consists of annotated samples from the medical field, enabling the transformation of $\mathcal{M}_{nat}$ into $\mathcal{M}_{med}$. Specifically, $\mathcal{D}_{med}$ is defined as a set of tuples $\{(x_i, c_i, \mathcal{S}_i)\}_{i=1}^{K}$, where K is the total number of image samples in the dataset. Each tuple includes a training image x_i, its corresponding image-level anomaly classification (AC) label $c_i \in \{-, +\}$, and pixel-level anomaly segmentation (AS) annotations $\mathcal{S}_i \in \{-, +\}^{h \times w}$ for images of size $h \times w$. The label '$+$' indicates an anomalous sample, while '$-$' denotes a normal one. For a given test image x_{test}, the model aims to accurately predict both image-level and pixel-level anomalies for AC and AS, respectively.

To model the detection of anomalies from unseen imaging modalities and anatomical regions, we approach the problem in a zero-shot learning context. Here, $\mathcal{D}_{med}$ is a pre-training dataset that is composed of medical data from different modalities and anatomical regions than those in the test samples, which assesses the model's generalization to unseen scenarios. Considering the practicality of obtaining a limited number of samples from the target scenario, we also extend the method to the few-shot learning context. Here, $\mathcal{D}_{med}$ includes a small collection of K annotated images that are of the same modality and anatomical region as those in the test samples, with K typically representing a small numerical value, such as $\{2, 4, 8, 16\}$ in this study.

Below we introduce our proposed multi-level adaptation and comparison framework for AD in medical images, comprising (i) multi-level feature adaptation (Sec. 4), and (ii) multi-level feature comparison (Sec. 5).

4. Train: Multi-Level Feature Adaptation

To adapt a pre-trained natural image visual-language model for anomaly detection (AD) in medical imaging, we introduce a multi-level feature adaptation framework specifically designed for AD in medical images, utilizing minimal data and lightweight multi-level feature adapters.

Multi-level Visual Feature Adapter (MVFA). Addressing the challenge of overfitting due to a high parameter count and limited training examples, we apply the CLIP adapter across multiple feature levels. This method appends a small set of learnable bottleneck linear layers to the visual branches of CLIP while keeping its original backbone unchanged, thus enabling adaptation at multiple feature levels.

As shown in Figure 2 (a), for an image $x \in \mathbb{R}^{h\times w\times 3}$, a CLIP visual encoder with four sequential stages (S_1 to S_4) transforms the image x into a feature space $\mathcal{F}_{vis} \in \mathbb{R}^{G\times d}$. Here, G represents the grid number, and d signifies the feature dimension. The output of the first three visual encoder stages (S_1 to S_3), denoted as $\mathcal{F}_l \in \mathbb{R}^{G\times d}, l \in \{1,2,3\}$, represents the three middle-stage features.

The visual feature adaptation involves three feature adapters, $A_l(\cdot)$, $l \in \{1,2,3\}$, and one feature projector, $P(\cdot)$, at different levels. At each level $l \in \{1,2,3\}$, a learnable feature adapter $A_l(\cdot)$ is integrated into the feature $\mathcal{F}_l$, encompassing two (the minimum number) layers of linear transformations. This integration transforms the features for adaptation, represented as:

$$A_l(\mathcal{F}_l) = ReLU(\mathcal{F}_l^T W_{l,1})W_{l,2}, \text{where } l \in \{1,2,3\}. \quad (1)$$

Here, $W_{l,1}$ and $W_{l,2}$ denote the learnable parameters of the linear transformations. Consistent with [15], a residual connection is employed in the feature adapter to retain the original knowledge encoded by the pre-trained CLIP. Specifically, a constant value γ serves as the residual ratio to adjust the degree of preserving the original knowledge for improved performance. Therefore, the feature adapter at the l-th feature level can be expressed as:

$$\mathcal{F}_l^* = \gamma A_l(\mathcal{F}_l)^T + (1-\gamma)\mathcal{F}_l, \text{where } l \in \{1,2,3\}, \quad (2)$$

with $\mathcal{F}_l^*$ serving as the input for the next encoder stage S_{l+1}. By default, we set $\gamma = 0.1$. Moreover, as shown in Figure 2 (b), to simultaneously address both global and local features for AC and AS respectively, a dual-adapter architecture replaces the single-adapter used in Eq. (1), producing two parallel sets of features at each level, $\mathcal{F}_{cls,l}$ and $\mathcal{F}_{seg,l}$. For the final visual feature $\mathcal{F}_{vis}$ generated by the CLIP visual encoder, a feature projector $P(\cdot)$ projects it using linear layers with parameters W_{cls} and W_{seg}, obtaining global and local features as $\mathcal{F}_{cls,4} = \mathcal{F}_{vis}^T W_{cls}$ and $\mathcal{F}_{seg,4} = \mathcal{F}_{vis}^T W_{seg}$. Utilizing the multi-level adapted features, the model is equipped to effectively discern both global anomalies for classification and local anomalies for segmentation, through the following visual-language feature alignment.

Language Feature Formatting. To develop an effective framework for anomaly classification and segmentation, we adopt a two-tiered approach for text prompts, inspired by methodologies used in [9, 26]. These methods leverage descriptions of both normal and abnormal objects. At the state level, our strategy involves using straightforward, generic text descriptions for normal and abnormal states, focusing on clarity and avoiding complex details. Moving to the template level, we conduct a thorough examination of the 35 templates referenced in [11] (detailed in the supplementary material). By calculating the average of the text features extracted by the text encoder for normal and abnormal states separately, we obtain a text feature represented as $\mathcal{F}_{text} \in \mathbb{R}^{2\times d}$, where d is the feature dimension.

Visual-Language Feature Alignment. For the image-level anomaly annotation $c \in \{-,+\}$ and the corresponding pixel-level anomaly map $\mathcal{S} \in \{-,+\}^{h\times w}$, we optimize the model at each feature level, $l \in \{1,2,3,4\}$, by aligning the adapted-visual features given by MVFA and the text features. This is achieved through a loss function that combines different components:

$$\begin{aligned}\mathcal{L}_l =& \lambda_1 \mathit{Dice}(\mathit{softmax}(\mathcal{F}_{seg,l}\mathcal{F}_{text}^T), \mathcal{S}) + \\ & \lambda_2 \mathit{Focal}(\mathit{softmax}(\mathcal{F}_{seg,l}\mathcal{F}_{text}^T), \mathcal{S}) + \\ & \lambda_3 \mathit{BCE}(\max_{h\times w}(\mathit{softmax}(\mathcal{F}_{cls,l}\mathcal{F}_{text}^T)), c),\end{aligned} \quad (3)$$

where $\mathit{Dice}(\cdot,\cdot)$, $\mathit{Focal}(\cdot,\cdot)$, and $\mathit{BCE}(\cdot,\cdot)$ are dice loss [36], focal loss [33], and binary cross-entropy loss, respectively. λ_1, λ_2 and λ_3 are the individual loss weights where we set $\lambda_1 = \lambda_2 = \lambda_3 = 1.0$ as default. The overall adaptation loss $\mathcal{L}_{adapt}$ is then calculated as the sum of losses at each feature level, expressed as $\mathcal{L}_{adapt} = \sum_{l=1}^{4} \mathcal{L}_l$.

Discussion. WinCLIP [26] relies on the class token from pre-trained VLMs for natural image AD, with no adaptation performed. In contrast, MVFA introduces multi-level adaptation, freezing the main backbone while adapting features at each level via adapters in line with corresponding visual-language alignments. The resulting adapted features are residually integrated into subsequent encoder blocks, modifying input features of these blocks. This unique approach enables collaborative training of adapters across different levels via gradient propagation, enhancing the overall adaptation of the backbone model. As a result, unlike APRIL-GAN [9], which utilizes isolated feature projections that do not adapt the main backbone, MVFA leads to robust generalization in medical AD. The difference between MVFA and feature projection in [9] is also evaluated in Sec. 6.3.

5. Test: Multi-Level Feature Comparison

During testing, to accurately predict anomalies at the image level (AC) and pixel level (AS), our approach incorporates a two-branch multi-level feature comparison architecture, comprising a zero-shot branch and a few-shot branch, as illustrated in Figure 2 (c).

Zero-Shot Branch. A test image x_{test} is processed through MVFA to produce multi-level adapted features. These features are then compared with the text feature $\mathcal{F}_{text}$. Zero-

Table 1. Comparisons with state-of-the-art **few-shot** anomaly detection methods with K=4. The AUCs (in %) for anomaly classification (AC) and anomaly segmentation (AS) are reported. The best result is in bold, and the second-best result is underlined.

Setting	Method	Source	HIS	ChestXray	OCT17	BrainMRI		LiverCT		RESC	
			AC	AC	AC	AC	AS	AC	AS	AC	AS
full-normal-shot	CFlowAD [19]	WACV 2022	54.54	71.44	85.43	73.97	93.52	49.93	92.78	74.43	93.75
	RD4AD [10]	CVPR 2022	66.59	67.53	97.24	89.38	96.54	60.02	95.86	87.53	96.17
	PatchCore [41]	CVPR 2022	69.34	75.17	98.56	91.55	96.97	60.40	96.58	91.50	96.39
	MKD [44]	CVPR 2022	77.74	81.99	96.62	81.38	89.54	60.39	96.14	88.97	86.60
few-normal-shot	CLIP [25]	OpenCLIP	63.48	70.74	98.59	74.31	93.44	56.74	97.20	84.54	95.03
	MedCLIP [52]	EMNLP 2022	75.89	**84.06**	81.39	76.87	90.91	60.65	94.45	66.58	88.98
	WinCLIP [26]	CVPR 2023	67.49	70.00	97.89	66.85	94.16	67.19	96.75	88.83	96.68
few-shot	DRA [12]	CVPR 2022	68.73	75.81	99.06	80.62	74.77	59.64	71.79	90.90	77.28
	BGAD [57]	CVPR 2023	-	-	-	83.56	92.68	72.48	98.88	86.22	93.84
	APRIL-GAN [9]	arXiv 2023	76.11	77.43	**99.41**	89.18	94.67	53.05	96.24	94.70	97.98
	MVFA	Ours	**82.71**	81.95	99.38	**92.44**	**97.30**	**81.18**	**99.73**	**96.18**	**98.97**

shot AC and AS results, denoted as c_{zero} and $\mathcal{S}_{\text{zero}}$, are calculated using average softmax scores across the four levels,

$$c_{\text{zero}} = \frac{1}{4}\sum_{l=1}^{4} \max_{G}(\textit{softmax}(\mathcal{F}_{cls,l}\mathcal{F}_{text}^{T})),$$
$$\mathcal{S}_{\text{zero}} = \frac{1}{4}\sum_{l=1}^{4} \text{BI}(\textit{softmax}(\mathcal{F}_{seg,l}\mathcal{F}_{text}^{T})). \quad (4)$$

Here, $\text{BI}(\cdot)$ reshapes the anomaly map to $\sqrt{G}\times\sqrt{G}$ and restores it to the original input image resolution using bilinear interpolation, with G representing the grid number.

Few-Shot Branch. All the multi-level visual features of a few labeled normal images in $\mathcal{D}_{med}$ contribute to constructing a multi-level feature memory bank $\mathcal{G}$ facilitating the feature comparison. The few-shot AC and AS scores, denoted as c_{few} and $\mathcal{S}_{\text{few}}$, are derived from the minimum distance between the test feature and the memory bank features at each level, through a nearest neighbor search process,

$$c_{\text{few}} = \frac{1}{4}\sum_{l=1}^{4} \max_{G}(\min_{m\in\mathcal{G}} \textit{Dist}(\mathcal{F}_{cls,l}, m)),$$
$$\mathcal{S}_{\text{few}} = \frac{1}{4}\sum_{l=1}^{4} \text{BI}(\min_{m\in\mathcal{G}} \textit{Dist}(\mathcal{F}_{seg,l}, m)). \quad (5)$$

Here, $\textit{Dist}(\cdot,\cdot)$ represents the cosine distance, calculated as $1-\text{cosine}(\cdot,\cdot)$. The final predicted AC and AS results combine the outcomes from both branches:

$$c_{\text{pred}} = \beta_1 c_{\text{zero}} + \beta_2 c_{\text{few}}, \mathcal{S}_{\text{pred}} = \beta_1 \mathcal{S}_{\text{zero}} + \beta_2 \mathcal{S}_{\text{few}}. \quad (6)$$

Here, β_1 and β_2 are weighting factors for the zero-shot and few-shot branches, respectively, set to 0.5 by default.

6. Experiments

6.1. Experimental Setups

Datasets. We consider a medical anomaly detection (AD) benchmark based on BMAD [3], covering five distinct med-

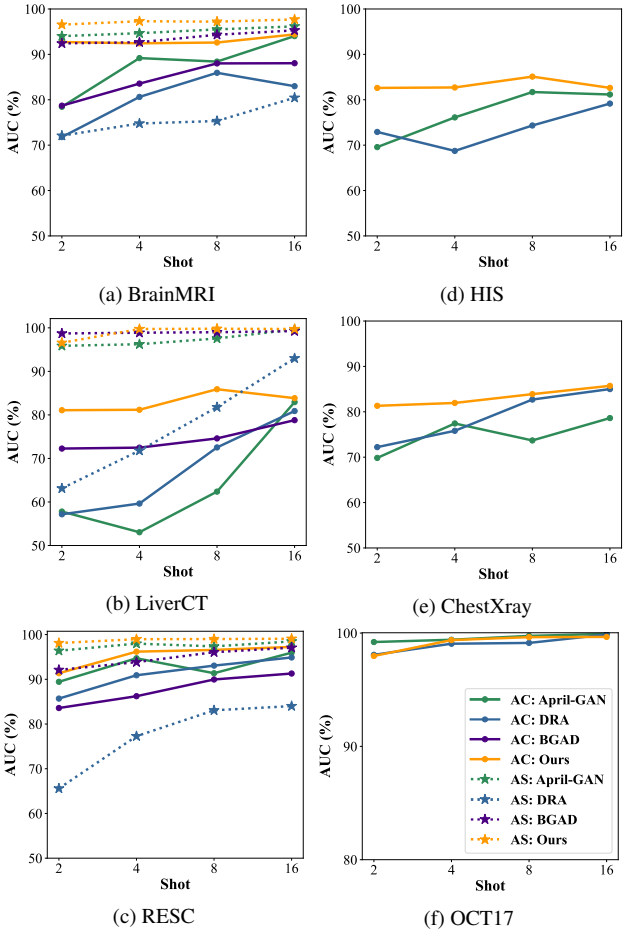


Figure 3. Comparisons with state-of-the-art **few-shot** anomaly detection methods on datasets of (a) BrainMRI, (b) LiverCT, (c) RESC, (d) HIS, (e) ChestXray, and (f) OCT17, with the shot number $K = \{2, 4, 8, 16\}$. The AUCs (in %) for anomaly classification (AC) and anomaly segmentation (AS) are reported. More details are included in the supplementary material.

ical domains and resulting in six datasets. These include brain MRI [1, 2, 35], liver CT [7, 29], retinal OCT [21, 28], chest X-ray [51], and digital histopathology [4]. Among these, BrainMRI [1, 2, 35], LiverCT [7, 29], and RESC [21] datasets are used for both anomaly classification (AC) and segmentation (AS), while OCT17 [28], ChestXray [51], and HIS [4] are solely for AC. Detailed descriptions of these datasets are provided in the supplementary material.

Competing Methods and Baselines. In this study, we consider various state-of-the-art AD methods within distinct training settings as competing methods. These settings encompass (i) vanilla methods that use all normal data (CFlowAD [19], RD4AD [10], PatchCore [41], and MKD [44]), (ii) few-normal-shot methods (CLIP [25], MedCLIP [52], WinCLIP [26]), and (iii) few-shot methods (DRA [12], BGAD [57], and April-GAN [9]). We evaluate these methods for AC and AS, excluding BGAD, which is exclusively applied for segmentation due to its requirement for pixel-level annotations during training.

Evaluation Protocols. The area under the Receiver Operating Characteristic curve metric (AUC) is used to quantify the performance. This metric is a standard in AD evaluation, with separate considerations for image-level AUC in AC and pixel-level AUC in AS.

Model Configuration and Training Details. We utilize the CLIP with ViT-L/14 architecture, with input images at a resolution of 240. The model comprises a total of 24 layers, which are divided into 4 stages, each encompassing 6 layers. We use the Adam optimizer at a constant learning rate of 1e-3 and a batch size of 16, conducting 50 epochs for training on one single NVIDIA GeForce RTX 3090 GPU.

6.2. Comparison with State-of-the-art Methods

Few-Shot Setting. In Table 1, we compare the performance of MVFA under the few-shot setting with $K = 4$ against other state-of-the-art AD methods. For an in-depth analysis of MVFA's performance across various few-shot scenarios ($K \in \{2, 4, 8, 16\}$), please refer to Figure 3. MVFA demonstrates superior performance over competing methods like DRA [12], BGAD [57], and April-GAN [9]. Notably, MVFA surpasses April-GAN, the winner of the VAND workshop at CVPR 2023 [68], by an average of 7.33% in AUC for AC and 2.37% in AUC for AS across all datasets. Compared to BGAD [57], MVFA shows an average improvement of 9.18% in AUC for AC and 3.53% in AUC for AS, in datasets with pixel-level annotations.

MVFA outperforms few-normal-shot CLIP-based methods such as CLIP [25] and WinCLIP [26], which also use visual-language pre-trained backbones and employ feature comparisons for AD. MVFA's advantage lies in its ability to effectively utilize a few abnormal samples, leading to significant gains over these methods. For example, against WinCLIP [26], MVFA achieves an average improvement of 12.60% in AUC for AC and 2.81% in AUC for AS across all datasets. While MedCLIP [52] shows superior results on ChestXray because it was trained on large-scale overlapping ChestXray data in our medical AD benchmark, it lacks broad generalization capabilities, as evidenced by its performance on other datasets.

Table 2. Comparisons with state-of-the-art **zero-shot** anomaly detection methods with in-/out-domain evaluation. The AUCs (in %) for AC and AS are reported.

Datasets	WinCLIP [26]	April-GAN [9]	MVFA (ours)
HIS	69.85 / -	72.36 / -	**77.90** / -
ChestXray	70.86 / -	57.49 / -	**71.11** / -
OCT17	46.64 / -	92.61 / -	**95.40** / -
BrainMRI	66.49 / 85.99	76.43 / **91.79**	**78.63** / 90.27
LiverCT	64.20 / 96.20	70.57 / 97.05	**76.24** / **97.85**
RESC	42.51 / 80.56	75.67 / 85.23	**83.31** / **92.05**

Table 3. Comparisons with state-of-the-art **few-shot** anomaly detection methods with K=4 for in-/out-domain evaluation. The AUCs (in %) for AC/AS are reported.

AC/AS (avg AUC%)	WinCLIP [26]	April-GAN [9]	MVFA
in-domain (MVTec)	95.16/96.27	92.77/95.89	**96.19/96.32**
out-domain (medical)	76.38/95.86	81.65/96.30	**88.97/98.67**

Moreover, MVFA exhibits substantial improvements over full-normal-shot vanilla AD methods such as CFlowAD [19], RD4AD [10], PatchCore [41], and MKD [44], which rely on much larger datasets than those employed in this study. This highlights the value of incorporating a few abnormal samples as supervision, especially in medical diagnostics where acquiring a limited number of abnormal data can be more practical.

Zero-Shot Setting. The experiments for zero-shot AD were conducted under the *leave-one-out setting*. In this configuration, a designated target dataset was chosen for testing, while the remaining datasets with different modalities and anatomical regions were employed for training. The aim of this approach was to gauge the performance when confronted with unseen modalities and anatomical regions, thereby assessing the model's capacity for generalization. Table 2 provides a comprehensive overview of the results pertaining to zero-shot medical AC and AS, offering a comparative evaluation alongside two state-of-the-art methods that also harness the capabilities of the CLIP backbone. MVFA shows remarkable superiority; for instance, it outperforms WinCLIP [26] with an average AUC improvement of 20.34% for AC and 5.81% for AS across all datasets. Similarly, against April-GAN [9], MVFA achieves an average AUC improvement of 6.24% for AC and 2.03% for AS across all datasets, underscoring its effectiveness in the challenging zero-shot medical AD setting.

In-Domain Evaluation. The MVTec AD benchmark [5],

Table 4. Ablation studies compared with feature alignment with multi-level projectors and feature adaptation with multi-level adapters under the zero-shot setting. The AUCs (in %) for AC and AS are reported.

Method	HIS	ChestXray	OCT17	BrainMRI		LiverCT		RESC	
	AC	AC	AC	AC	AS	AC	AS	AC	AS
feature alignment (projectors)	66.32	58.06	49.85	76.66	89.39	75.85	97.64	74.44	89.17
feature adaptation (adapters)	**77.90**	**71.11**	**95.40**	**78.63**	**90.27**	**76.24**	**97.85**	**83.31**	**92.05**

Table 5. Ablation studies of multi-level feature ensemble under *single-level training/multi-level training* with the same model architecture. The average AUCs (in %) of all six datasets for AC and AS under the few-shot setting (K=4) are reported. Results for each dataset are included in the supplementary material.

	Layer 1	Layer 2	Layer 3	Layer 4	All
AC	80.96/83.39	88.83/88.84	81.25/84.32	83.33/84.62	**88.97**
AS	97.70/97.98	98.58/98.62	96.03/98.54	97.19/97.44	**98.67**

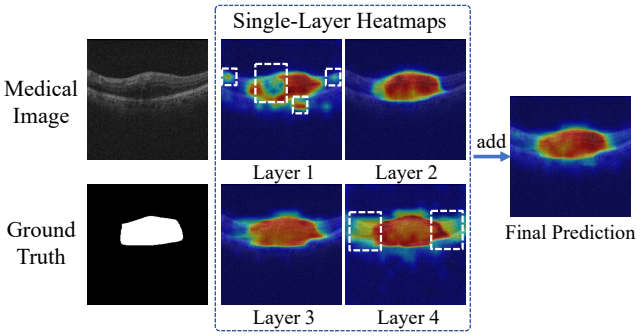


Figure 4. Considering features across multiple levels significantly enhances segmentation performance. The white dashed boxes demarcate regions that have been missed or erroneously segmented. More cases are included in the supplementary material.

consisting of 15 industrial defect detection sub-datasets, is considered as in-domain evaluation. As shown in Table 3, although our main focus is not in-domain scenarios, MVFA shows comparable performance to state-of-the-art methods in a few-shot setting (K=4). Detailed results for each sub-dataset are included in the supplementary material. In our main focus of out-domain evaluations on the medical AD benchmark, MVFA significantly outperforms competing methods, highlighting its superior generalization capabilities.

6.3. Ablation Studies

Feature Adaptation vs. Feature Alignment. We conduct ablation studies in the zero-shot setting for AC and AS to assess the effectiveness of multi-level feature adapters in enhancing cross-modal generalization. For this purpose, we substitute the multi-level feature adapters with distinct multi-level feature projectors, while keeping the parameters and feature alignment loss functions the same. The primary distinction is that each projector is optimized independently, in contrast to the collective optimization approach used for adapters. This ensures that the only variable under consideration is the architecture, while all other factors, including model parameters and training loss functions, remain constant for a valid comparison.

The results, as shown in Table 4, reveal significant improvements with feature adapters. They lead to substantial improvements in AC, with image-level AUC increasing by 11.58%, 13.05%, 45.55%, 1.97%, 0.39%, and 8.87% for HIS, ChestXay, OCT17, BrainMRI, LiverCT, and RESC, respectively, with an average improvement of 13.57%. For AS, improvements were observed, with pixel-level AUC increasing by 0.88%, 0.21%, and 2.88% for BrainMRI, LiverCT, and RESC, respectively. These findings highlight the critical role of multi-level feature adapters in boosting the model's generalization capabilities.

Feature Ensemble in Multi-Level. In this study, as shown in Table 5, we evaluate the effectiveness of an ensemble approach that integrates features from different layers. The approaches are compared against our comprehensive model, both using the multi-level adapter architecture. Our evaluation specifically focused on features from four distinct layers under two training scenarios: (i) single-layer training, where only one layer's adapter is optimized in each experiment, and (ii) multi-level training, aligning with the methodology of our comprehensive model. The objective of this comparison was to determine the benefits of combining features from multiple layers compared to optimizing each layer's features separately.

The results presented in Table 5 demonstrate that among individual layers, Layer 2 yielded the highest performance, achieving 88.84% in AC and 98.62% in AS. However, the ensemble method, which integrates features from all layers, outperformed single-layer approaches, recording AUCs of 88.97% in AC and 98.67% in AS. This underscores the effectiveness of combining features from multiple levels. Moreover, multi-level training consistently exceeds the performance of single-layer training, reinforcing the benefits of our multi-level adaptation approach across all layers.

The visualizations of retina anomaly segmentation from various layers, as depicted in Figure 4, further reinforce our

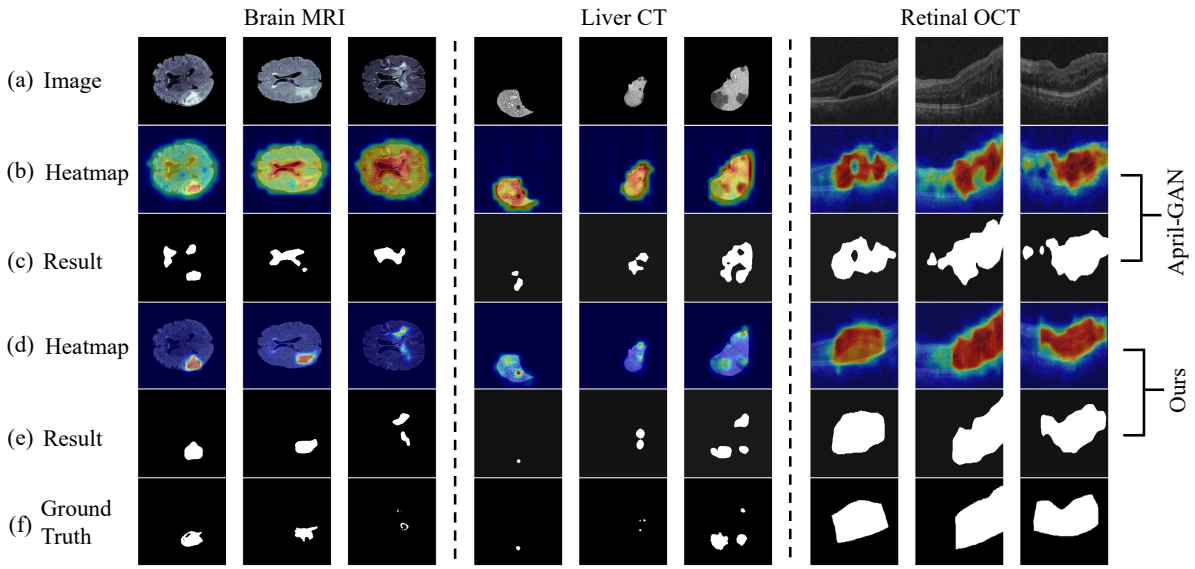
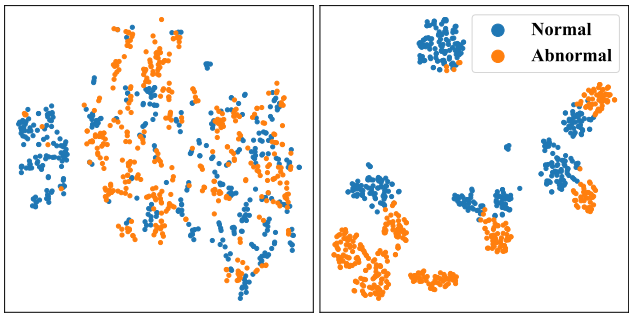


Figure 5. Visualization of AS for MVFA on brain MRI, liver CT, and retinal OCT, compared with state-of-the-art method April-GAN. Results from (e) show better performance than results from (c), showing the effectiveness of the proposed multi-level feature adaptation.



(a) Without Adapter (b) With Adapter

Figure 6. Visualization, using t-SNE, of the features learned from the RESC test set, using (a) pretrained visual encoder, and (b) multi-level feature adapters. The same t-SNE optimization iterations are used in each case. Results show that features extracted by adapters are separated between normal and abnormal samples.

findings. These visualizations distinctly show that single-layer methods are less effective compared to the ensemble approaches. By synergistically integrating features from different layers, we achieve a marked improvement in AD, which aligns with and supports our quantitative results.

6.4. Visualization Analysis

To qualitatively analyze how the proposed multi-level feature adaptation approach improves anomaly segmentation performance, we visualize the results of several cases from the BrainMRI, LiverCT, and RESC datasets. It can be seen from the results in Figure 5 that the segmentation produced by our MVFA (column e) is closer to the ground truth (column f) than that produced by the state-of-the-art method April-GAN (column c). This illustrates the effectiveness of the proposed multi-level visual feature adaptation.

We employ t-SNE [49] to visualize the learned features from RESC, as depicted in Figure 6. Each dot corresponds to features of a normal or abnormal sample from the test set. It can be seen that adapter enhances the separation between samples belonging to distinct states, which is beneficial for the identification of AD decision boundaries.

7. Conclusion

This paper adapts the pretrained visual-language models in natural domain to medical AD, with cross-domain generalizability among various modalities and anatomical regions. The adaptation involves not only from natural domain to medical domain, but also from high-level semantics to pixel-level segmentation. To achieve such goals, a collaborative multi-level feature adaptation method is introduced, where each adaptation is guided by the corresponding visual-language alignments, facilitating segmenting anomalies of diverse forms from medical images. Coupled with a comparison-based AD strategy, the method enables flexible adaptation for datasets with substantial modality and distribution differences. The proposed method outperforms existing methods on zero-/few-shot AC and AS tasks, indicating promising research avenues for future exploration.

Acknowledgment. This work is supported by the National Key R&D Program of China (No. 2022ZD0160703), STCSM (No. 22511106101, No. 18DZ2270700, No. 21DZ1100100), 111 plan (No. BP0719010), State Key Laboratory of UHD Video and Audio Production and Presentation, and the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award Number: MOE-T2EP20122-0006).

References

[1] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 2, 6

[2] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data*, 4(1): 1–13, 2017. 2, 6

[3] Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection. *arXiv preprint arXiv:2306.11876*, 2023. 2, 5

[4] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. 2, 6

[5] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 2019. 2, 6

[6] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *CVPR*, 2020. 2

[7] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023. 2, 6

[8] Yu Cai, Hao Chen, Xin Yang, Yu Zhou, and Kwang-Ting Cheng. Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images. *Medical Image Analysis*, 86:102794, 2023. 2

[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. 1, 2, 4, 5, 6

[10] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *CVPR*, 2022. 2, 5, 6

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4

[12] Choubo Ding, Guansong Pang, and Chunhua Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *CVPR*, 2022. 1, 2, 5, 6

[13] Zhiyuan Ding, Qi Dong, Haote Xu, Chenxin Li, Xinghao Ding, and Yue Huang. Unsupervised anomaly segmentation for brain lesions using dual semantic-manifold reconstruction. In *International Conference on Neural Information Processing*, 2022. 2

[14] 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

[15] 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. 4

[16] 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

[17] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *ICCV*, 2019. 2

[18] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2021. 2

[19] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *WACV*, 2022. 2, 5, 6

[20] Keji He, Chenyang Si, Zhihe Lu, Yan Huang, Liang Wang, and Xinchao Wang. Frequency-enhanced data augmentation for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, 2023. 2

[21] Junjie Hu, Yuanyuan Chen, and Zhang Yi. Automated segmentation of macular edema in oct using deep neural networks. *Medical Image Analysis*, 55:216–227, 2019. 2, 6

[22] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *ECCV*, 2022. 1

[23] Chaoqin Huang, Aofan Jiang, Ya Zhang, and Yanfeng Wang. Multi-scale memory comparison for zero-/few-shot anomaly detection. *arXiv preprint arXiv:2308.04789*, 2023. 2

[24] Weikai Huang, Yijin Huang, and Xiaoying Tang. Lesion-paste: One-shot anomaly detection for medical images. *arXiv preprint arXiv:2203.06354*, 2022. 2

[25] 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. 2, 5, 6

[26] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, 2023. 1, 2, 4, 5, 6

[27] Aofan Jiang, Chaoqin Huang, Qing Cao, Shuang Wu, Zi Zeng, Kang Chen, Ya Zhang, and Yanfeng Wang. Multi-scale cross-restoration framework for electrocardiogram anomaly detection. In *MICCAI*, 2023. 2

[28] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018. 2, 6

[29] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 2, 6

[30] Chenxin Li, Yunlong Zhang, Jiongcheng Li, Yue Huang, and Xinghao Ding. Unsupervised anomaly segmentation using image-semantic cycle translation. *arXiv preprint arXiv:2103.09094*, 2021. 2

[31] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *CVPR*, 2021. 2

[32] Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. In *Advances in Neural Information Processing Systems*, 2023. 2

[33] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 4

[34] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *IEEE/CVF Conference Computer Vision and Pattern Recognition*, 2024. 2

[35] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014. 2, 6

[36] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016. 4

[37] Guansong Pang, Chunhua Shen, and Anton van den Hengel. Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 353–362, 2019. 2

[38] 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 *ICML*, 2021. 1, 2

[39] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, 2022. 2

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[41] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, 2022. 2, 5, 6

[42] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *ICLR*, 2020. 2

[43] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *CVPR*, 2018. 2

[44] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad Hossein Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *CVPR*, 2021. 2, 5, 6

[45] 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. *NeurIPS*, 2022. 2

[46] Shelly Sheynin, Sagie Benaim, and Lior Wolf. A hierarchical transformation-discriminating generative model for few shot anomaly detection. In *ICCV*, 2021. 1

[47] Jianpo Su, Hui Shen, Limin Peng, and Dewen Hu. Few-shot domain-adaptive anomaly detection for cross-site brain images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (Early Access)*, 2021. 1

[48] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *NeurIPS*, 2020. 2

[49] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (11), 2008. 8

[50] Shenzhi Wang, Liwei Wu, Lei Cui, and Yujun Shen. Glancing at the patch: Anomaly localization with global and local feature comparison. In *CVPR*, 2021. 2

[51] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017. 2, 6

[52] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *EMNLP*, 2022. 5, 6

[53] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. In *MICCAI*, 2022. 2

[54] Jhih-Ciang Wu, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Learning unsupervised metaformer for anomaly detection. In *ICCV*, 2021. 2

[55] Haote Xu, Yunlong Zhang, Liyan Sun, Chenxin Li, Yue Huang, and Xinghao Ding. Afsc: Adaptive fourier space compression for anomaly detection. *arXiv preprint arXiv:2204.07963*, 2022. 2

[56] Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *IEEE International Conference on Computer Vision*, 2023. 2

[57] Xincheng Yao, Ruoqi Li, Jing Zhang, Jun Sun, and Chongyang Zhang. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *CVPR*, 2023. 1, 2, 5, 6

[58] Fei Ye, Chaoqin Huang, Jinkun Cao, Maosen Li, Ya Zhang, and Cewu Lu. Attribute restoration framework for anomaly detection. *IEEE Transactions on Multimedia*, 24:116–127, 2022. 2

[59] Jingwen Ye, Ruonan Yu, Songhua Liu, and Xinchao Wang. Mutual-modality adversarial attack with semantic perturbation. In *AAAI Conference on Artificial Intelligence*, 2024. 2

[60] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *IEEE/CVF Conference Computer Vision and Pattern Recognition*, 2023. 2

[61] Jianpeng Zhang, Yutong Xie, Zhibin Liao, Guansong Pang, Johan Verjans, Wenxin Li, Zongji Sun, Jian He, and Chunhua Shen Yi Li. Viral pneumonia screening on chest x-ray images using confidence-aware anomaly detection. *IEEE Transactions on Medical Imaging*, 40(3):879–890, 2021. 1, 2

[62] Ruipeng Zhang, Ziqing Fan, Qinwei Xu, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Grace: A generalized and personalized federated learning method for medical imaging. In *MICCAI*, 2023. 2

[63] 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 *CVPR*, pages 16793–16803, 2022. 2

[64] Kang Zhou, Yuting Xiao, Jianlong Yang, Jun Cheng, Wen Liu, Weixin Luo, Zaiwang Gu, Jiang Liu, and Shenghua Gao. Encoding structure-texture relation with p-net for anomaly detection in retinal images. In *ECCV*, 2020. 2

[65] Kang Zhou, Jing Li, Weixin Luo, Zhengxin Li, Jianlong Yang, Huazhu Fu, Jun Cheng, Jiang Liu, and Shenghua Gao. Proxy-bridged image reconstruction network for anomaly detection in medical images. *IEEE Transactions on Medical Imaging*, 41(3):582–594, 2021.

[66] Kang Zhou, Jing Li, Yuting Xiao, Jianlong Yang, Jun Cheng, Wen Liu, Weixin Luo, Jiang Liu, and Shenghua Gao. Memorizing structure-texture correspondence for image anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2335–2349, 2021. 2

[67] 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. 2

[68] Yang Zou, Taewan Kim, Latha Pemula, and Onkar Dabeer. Visual anomaly and novelty detection (vand) challenge in cvpr 2023 workshop, 2023. https://sites.google.com/view/vand-cvpr23/challenge. 6