Abnormality Detection In Medical Image Based On Visual-Language Model

Hoang-Phu Thanh-Luong*[1,2], Van-Thai Vu*[1,2], and Quoc-Ngoc Ly[1,2]

[1] University of Science, Ho Chi Minh City, Vietnam
[2] Viet Nam National University, Ho Chi Minh City, Vietnam
{20120548,20120579}@student.hcmus.edu.vn, lqngoc@fit.hcmus.edu.vn

Abstract. The application of artificial intelligence in medicine is receiving increasing attention and research. One of the critical challenges faced by doctors is detecting abnormalities in medical images. For this task, doctors expect a model not only to identify whether an image is normal or abnormal but also to accurately localize the areas of abnormalities. Additionally, the model needs to perform well across various medical imaging domains, even in fields with limited training data. Building models to address this issue typically requires large datasets and considerable time, which increases costs. However, the development of vision-language models, which can effectively handle zero-shot and few-shot problems, presents a more optimal solution. In this article, we propose a **M**ulti-**L**evel **A**dapters with **L**earnable **P**rompt model (MLA-LP). By adjusting image characteristics through adapters, our model can better calibrate centroids for medical datasets. And it automatically adjusts prompts to enhance its ability to detect abnormalities in medical images. Our experiments on medical anomaly detection benchmarks demonstrate that our method significantly surpasses current state-of-the-art models. In few-shot settings, our model achieves an average AUC improvement of 0.28% for anomaly classification and 0.08% for anomaly segmentation. In zero-shot settings, it achieves an average AUC improvement of 0.39% for anomaly classification.

Keywords: Anomaly detection · Vision-Languages Model · Medical image· Zero-shot learning · Few-shot learning.

1 Introduction

The medical anomaly detection (AD) problem involves determining whether an image is normal or abnormal and identifying abnormal areas on the medical image. This helps doctors quickly screen patients, reduces the risks in decision-making, and improves the efficiency of medical professionals.

Medical imaging varies widely, creating a need for a model that works well with different medical data sets. Some methods achieve a good understanding of the problem and perform almost accurately on certain tasks. However, these models require a large amount of data, increasing the cost and time to build them. Zero- or few-shot learning methods address this issue and also achieve

good results. With only a small amount of training data needed, these approaches offer new solutions for detecting abnormalities in medical images.

Recently, the pre-trained visual-language models (VLMs) have been increasingly improved on a larger scale. This greatly supports the problem of anomaly detection. One of the prominent models is CLIP [19], with the capability to map natural images and raw texts into a unified representational space, which can be easily applied to a variety of downstream tasks. This provides good support for many tasks, including anomaly detection. Several studies by WinCLIP [13] or April-GAN [5] have proposed different approaches based on the general idea of adding adapter layers to map image features to the joint embedding space to the text features, facilitating their comparison. And it has also produced positive results. In medicine, MVFA [10]'s proposal is also based on the above idea and has also achieved very good results for detecting abnormalities in medicine.

In this paper, we aim to develop a model for medical anomaly detection that is adaptable to zero-shot and few-shot learning methods. However, we face two main challenges when using the CLIP [19] model with adjustments. First, the visual encoder in CLIP [19] primarily represents semantic features of images, while our model needs to detect irregularities across diverse semantic contexts. Additionally, since CLIP [19] is trained only on natural images, applying it to medical images presents a challenge. Second, current anomaly detection methods often rely on additional information such as text data and pre-defined sentence templates, which reduces the model's flexibility in adapting to different datasets from various body parts. This is especially problematic when aiming for effective performance in few-shot and zero-shot learning scenarios.

To address these challenges, we propose a **M**ulti-**L**evel **A**dapters with **L**earnable **P**rompt model (MLA-LP). This model uses multi-level adaptation to align intermediate layer features from CLIP [19], allowing for adaptability at multiple levels and improving performance in medical anomaly detection. At the same time, we use learnable prompts to avoid static text templates, increasing the flexibility of the textual component in the anomaly detection task. Our goal is to recalibrate the model to classify images as 'normal' or 'abnormal' and to segment abnormalities within the images.

We experiment with MLA-PL using benchmark datasets for medical anomaly detection (AD). This includes five datasets with distinct medical modalities and anatomical regions: ChestXray [20], HIS [3], OCT17 [14], BrainMRI [1, 2, 17], LiverCT [4, 15], and RESC [9]. We compare the results with those of current state-of-the-art models [10]. Our approach shows superior performance across multiple datasets in both zero-shot and few-shot learning scenarios.

2 Related Works

Medical Anomaly Detection. In the application of Vision-Language Models to specific tasks in general domains and medical fields in particular, there are two prominent approaches. The first involves developing large-scale models, such as MedCLIP [21] and GLoRIA [12], and utilizing their pretrained encoders for downstream tasks like classification or segmentation,... The second

approach, which aligns with our work, has been explored in studies such as WinCLIP [13], April-GAN [5], and MVFA-AD [10], the current state-of-the-art for medical anomaly detection. These methods use CLIP [19] as the backbone, incorporating task-specific adaptations to directly address medical anomaly detection. This approach leverages CLIP's capacity for multimodal alignment while introducing enhancements tailored to the unique challenges of medical imaging.

Vision-Language Model. CLIP [19] (Contrastive Language-Image Pre-training) has shown strong potential in anomaly detection [5, 13, 25], including medical anomaly detection [10, 23, 25], thanks to its training on over 400 million image-text pairs, enabling effective visual-textual alignment. This capability aids in detecting rare pathologies with minimal labeled data, addressing challenges such as dataset scarcity and variability in medical imaging. Huang et al. [10] introduced multi-level adaptation to the image encoder of CLIP [19] by freezing both the image and text encoders while training lightweight adapters. These adapters successfully adapt medical features from medical images, achieving strong performance in specialized tasks. Zhou et al. [25] enhanced CLIP by adding DPAM layers to the image encoder and applying textual prompt tuning to the text encoder, enabling prompt learning in both textual and local visual spaces. This approach was inspired by the prompt tuning methodology proposed in Zhou et al. [24]. Similarly, Zhang et al. [23] improved the text encoder by incorporating learnable prompts and integrating multi-task anomaly synthesis tailored specifically for medical anomaly detection. Inspired by Huang et al.'s [10] approach of adding multi-level adaptation to the image encoder of CLIP and the use of learnable prompts tailored for specific tasks, as demonstrated in the works of Zhou et al. [25] and Zhang et al. [23], we aim to apply these techniques to develop an architecture for the task of Abnormality Detection in Medical Images based on a Visual-Language Model.

3 Methods

In this section, we detail our MLA-PL method, starting with the use of learnable prompts in text encoding, followed by multi-level adaptation in image encoding to capture detailed medical features. We then describe the integration of image and text features for improved abnormality detection.

3.1 Learnable Prompt

In medical anomaly detection, leveraging learnable prompts in text-based encoding provides a powerful mechanism for improving system efficiency. Text inherently carries rich semantic information and offers an abstract, flexible means to represent abnormalities in medical imaging. This enables the system to dynamically adapt to various contexts, enhancing feature extraction and improving the performance of classification and segmentation tasks.

The main idea behind learnable prompts is to model a prompt's context using a set of **learnable vectors**, which are optimized by minimizing the classification loss and segmentation loss. Instead of relying solely on handcrafted templates, learnable prompts allow the model to automatically adjust and generalize across

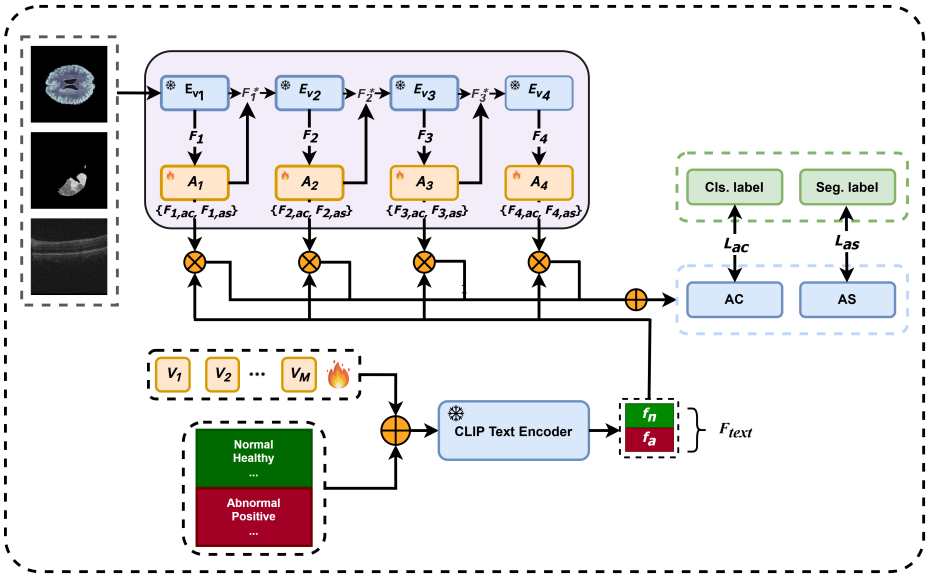

Fig. 1: The overview of our proposed MLA-LP model.

diverse medical imaging scenarios, capturing nuanced features of normal and abnormal cases effectively.

The commonly used text prompt templates in CLIP [19], such as 'A photo of a [CLS],' primarily focus on describing the semantics of the image. Therefore, when used for anomaly detection (AD), it is necessary to include information about 'normal' or 'abnormal,' for example, 'A photo of a [CLS] with abnormal.' However, this also requires building many sentence patterns and information about CLS like brain, chest,*etc*., which poses additional challenges for this anomaly detection problem. Therefore, we use prompt learning instead of the above traditional method, which will generalize better to medical images to obtain more comprehensive abnormal semantics. Going into more detail, we use templates as:

$$p = [V_1][V_2]\cdots[V_M][\texttt{CLS}] \tag{1}$$

Where V_i $(i \in [1, \ldots, M])$ represents the learned embeddings, M is the number of learnable tokens. And the class token (CLS) is the fixed embeddings, Therefore, we use medical terms to describe the two normal and abnormal cases for [CLS]. For example, words like [healthy], [asymptomatic] used for normal cases, and [disease], [pathological] for abnormal cases. In our study, we use a set of sentences $P_n = \{p_{n_1}, p_{n_2}, \ldots, p_{n_I}\}$ for normal cases and $P_a = \{p_{a_1}, p_{a_2}, \ldots, p_{a_E}\}$ for abnormal cases. Therein, I, E are the number of prompts of normal and abnormal.

The text encoder of CLIP is denoted as $E_{text}(.)$. For each sentence in P_n, P_a, it pass through $E_{text}(.)$. From there, we get $F_n = \{f_{n_1}, f_{n_2}, \ldots, f_{n_I}\}$ and $F_a = \{f_{a_1}, f_{a_2}, \ldots, f_{a_E}\}$. For each feature $f_{n_i}, f_{a_j} \in R^C$, where C signifies the dimension of prompt feature.

Finally, we perform the mean operation on each F_n and F_a to obtain their respective aggregated features:

$$f_n = \frac{1}{I}\sum_{i=1}^{I} f_{n_i}, \;\; f_a = \frac{1}{E}\sum_{i=1}^{E} f_{a_i} \tag{2}$$

Combining these results from Equation 2, the final text feature F_{text} is represented as:

$$F_{text} = \{f_n, f_a\} = \begin{pmatrix} f_n^T \\ f_a^T \end{pmatrix} \tag{3}$$

Since f_n and f_a are both feature embeddings in the space $\mathbb{R}^C$, will be a matrix of size $2 \times C$. The text feature will be used for integration with image features.

By integrating these textual features with multi-level image adaptations, detailed medical features are captured and fused, leading to superior performance in detecting and segmenting abnormal regions in medical images.

3.2 Multi-Level Adaptation

In this section, we will describe how we integrate the adapters with the CLIP [19]'s image encoder using the multi-level adaptation mechanism. With a small number of learnable parameters, this can help us address the AD problem effectively. We will present three parts: CLIP [19]'s image encoder, adapter architecture, and integration of CLIP [19]'s image encoder and the adapters.

3.2.1 CLIP image encoder

With the approach of using the ViT model [7] for the CLIP [19]'s image encoder, as shown in Fig. 1, an image $I \in \mathbb{R}^{H\times W\times 3}$, where H and W denote the height and width of the image. Firstly, It changed into sequence of 2D patches $I_p \in \mathbb{R}^{N_p\times S^2\times 3}$, where S and N_p represent the patch size and the number of patches. Then, the transformer layer projects I_p into the embedded feature space. Subsequently, it passes through multiple transformer layers. We denote the feature at the i-th stage as $F_i \in \mathbb{R}^{N_p\times D}$, where D denotes the embedding dimension.

3.2.2 Adapter architecture

In this section, we present the converter used to extract local features to support the two problems of classification and segmentation. By combining linear layers together according to a bottleneck design [8].

From the Fig. 2 (a), the input to the adapter is F, a feature extracted from the CLIP model [19]. Each adapter is composed of two components, A_{ac} and A_{as}, responsible for processing classification and segmentation tasks, respectively. Both A_{ac} and A_{as} adopt a bottleneck architecture [8], enabling efficient utilization of extracted features before passing them back to the CLIP image encoder [19]. From the Fig. 2 (b), the detailed architecture of the adapter consists of two linear layers with a LeakyReLU activation layer in between.

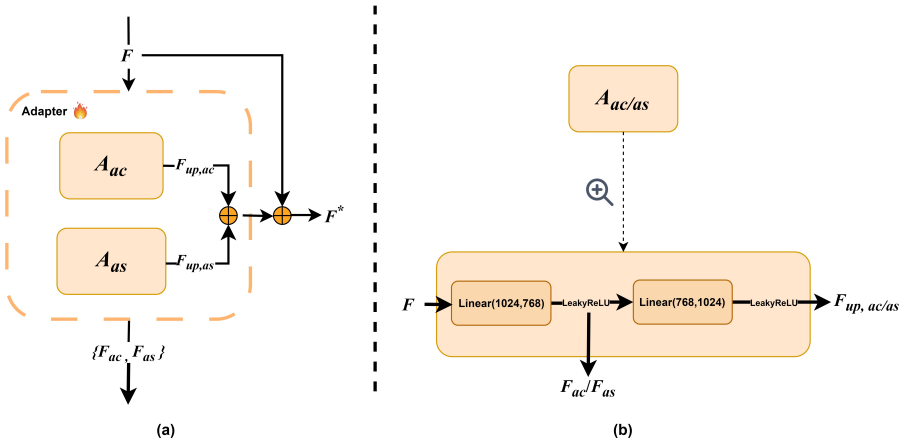

Fig. 2: Illustration of the Visual Adapter architecture. (a) The left figure presents the general architecture of the Visual Adapter, comprising two modules: the classification adapter A_{ac} and the segmentation adapter A_{as}. (b) The right figure details the architecture of each adapter module $A_{ac/as}$.

The adapter will produce $\{F_{ac}, F_{as}\}$ which will be combined with text feature F_{text} later, and $\{F_{up,ac}, F_{up,as}\}$ which will be integrated back with the input feature F_f. Thus, each adapter receives F'_f as input and will go through two parts A_{ac} and A_{as}. The output of each adapter includes:

- F^* is the feature that will be obtained by combining the features from $F_{up,ac}$, $F_{up,as}$ and F, calculated by formula:

$$F^* = \gamma F + \alpha F_{up,ac} + \beta F_{up,as} \tag{4}$$

  With γ, α, β being the ratio to adjust the level of initial knowledge retention to improve the efficiency of the model, the team is using a parameter set of $\gamma = 0.8, \alpha = 0.1, \beta = 0.1$.
- F_{ac}, F_{as} are features used for classification and segmentation problems.

3.2.3 Integration of CLIP image encoder and adapters

In the CLIP image encoder [19] with the ViT [7] architecture, after the image undergoes several layers to become I_p, the sequence of transformed classes will be divided into four sequential stages (S_1 to S_4), with an integration transformation module A_l with $l \in \{1, 2, 3, 4\}$ between each stage.

At each transformation module A_l , $l \in \{1, 2, 3, 4\}$, the input will be the feature $F_{l,f}$ obtained at each stage $l \in \{1, 2, 3, 4\}$. The output includes $F_{l,ac}$, $F_{l,as}$, which is used to calculate with text features, and F^*_l, which is used as input for the next stage.

3.3 Integration of features between images and text

After presenting how to obtain the text feature F_{text} and the pairs of image features $\{F_{l,ac}, F_{l,as}\}$ with $l \in \{1, 2, 3, 4\}$, in this section, we will describe

how we combine them to calculate for both anomaly classification and anomaly segmentation tasks.

3.3.1 Anomaly Classification

For the classification task, we calculate the cosine similarity between the adapter modules $F_{l,ac}, l \in \{1,2,3,4\}$ and the text feature F_{text}. Thus, for each $F_{l,ac}$, we obtain:

$$P_{l,ac} = Norm(\texttt{softmax}(F_{l,ac} * F_{text})), l \in \{1,2,3,4\} \tag{5}$$

When performing matrix multiplication, we apply both the $\texttt{softmax}(.)$ and normalization $Norm(.)$ operations to process the outputs.

For each layer $l \in \{1,2,3,4\}$, the predictions $P_{l,ac}$ consist of two components:

$$P_{l,ac} = \{P^n_{l,ac}, P^a_{l,ac}\}, \tag{6}$$

where $P^n_{l,ac}$ and $P^a_{l,ac}$ represent the predictions for the normal and abnormal classification labels, respectively.

Here, $S_{ac} \in \{+,-\}$ denotes the anomaly classification label, $+$ represents abnormal images, and $-$ represents normal images. To optimize the classification task, we use the binary cross-entropy loss function:

$$L_{ac} = \sum_{l=1}^{4} L_{bce}(P_{l,ac}, S_{ac}), l \in \{1,2,3,4\} \tag{7}$$

Finally, we obtain the classification results by summing the outputs of $P^a_{l,ac}, l \in \{1,2,3,4\}$

$$P_{ac} = \sum_{i=1}^{n=4} P^a_{l,ac} \tag{8}$$

3.3.2 Anomaly Segmentation

For the segmentation task, we calculate the cosine similarity between the adapter modules $F_{l,as}, l \in \{1,2,3,4\}$ and the text feature F_{text}, derived from learnable prompts. Thus, for each $F_{l,as}$, we obtain:

$$P'_{l,as} = \texttt{softmax}(F_{l,as} * F_{text}), l \in \{1,2,3,4\} \tag{9}$$

Subsequently, we perform operation $BI(.)$, Bilinear Interpolation, to shape the anomaly map into $S \times S$ and resize it back to the original dimensions of the input image using bilinear interpolation:

$$P_{l,as} = BI(P'_{l,as}), l \in \{1,2,3,4\} \tag{10}$$

After the above calculations, we obtain four segmentation labels $P_{l,as}$, $l \in \{1,2,3,4\}$ with each $P_{l,as} = \{P^n_{l,as}, P^a_{l,as}\}$ denotes the normal and abnormal segmentation labels.

To optimize, $S_{seg} \in \mathbb{R}^{H \times W}$ denote segmentation label, and S_{seg} is a matrix with 2 values: 0 and 1. We use Focal [16] and Dice [18] to optimize the segmentation task. The loss function Focal [16] is particularly effective in addressing class imbalance issues.

$$\begin{aligned} L_{as,l} = & L_{focal}(P_{l,as}, S_{as}) + L_{dice}(P^a_{l,as}, S_{as}) \\ & + L_{dice}(P^n_{l,as}, 1 - S_{as}),\ l \in \{1, 2, 3, 4\} \end{aligned} \tag{11}$$

Finally, to obtain the final segmentation label, we perform matrix addition on the four values $P_{l,as}$ obtained above:

$$P_{as} = \sum_{i=1}^{n=4} P_{l,as} \tag{12}$$

The segmentation label P_{as} is thus created by aligning the text-derived semantics with multi-level visual features, effectively capturing both normal and abnormal regions in medical images. Optimizing this process with Focal Loss [16] and Dice Loss [18] ensures accurate and robust segmentation.

4 Experiments

4.1 Experimental Setup

Datasets. We conducted experiments on six distinct medical datasets spanning various domains: brain MRI [1, 2, 17], liver CT [4, 15], retinal OCT [9, 14], chest X-ray [20], and digital histopathology (HIS) [3]. The Brain MRI dataset [1, 2, 17] comprises 2D brain MRI images, including both normal and abnormal cases (affected by cancer). The Liver CT dataset [4, 15] is constructed from two datasets: BTCV [4] and LiTS [15]. BTCV includes 50 3D abdominal CT scans, while LiTS consists of 131 3D abdominal CT scans. The Retinal OCT dataset [9, 14] contains two distinct OCT datasets. The RESC dataset [9] provides segmentation labels, delineating areas affected by macular edema, whereas the OCT17 dataset [14] is intended for classification tasks, containing retinal OCT images categorized into three types of abnormalities. The Chest X-ray dataset [20] comprises 108,948 frontal-view X-ray images of 32,717 unique patients, annotated with eight disease labels obtained through text mining. Lastly, the HIS dataset [3] includes 400 whole slide images (WSI) of lymph node sections stained with hematoxylin and eosin (H&E) from breast cancer patients.

Baselines and Metrics. To comprehensively evaluate our proposed model, we compare it with prior state-of-the-art (SOTA) models. For the zero-shot approach, we compare our model with three other models: WinCLIP [13], April-GAN [5], and MVFA-AD [10]. For the few-shot approach, we compare it with four methods: DRA [6], BGAD [22], April-GAN [5], and MVFA-AD [10]. Experimental results for WinCLIP [13], April-GAN [5], MVFA-AD [10], and DRA [6], BGAD [22] are obtained from the study by [11]. We utilize the area under the Receiver Operating Characteristic curve metric (AUC) as the evaluation metric. This metric serves as a standard measure for anomaly detection, divided into two

Table 1: The results are compared with state-of-the-art models in medical image anomaly detection. The outcomes are evaluated using the AUC (Area Under the Curve) metric (%) for both anomaly classification (AC) and anomaly segmentation (AS). The performance of previous methods was collected from the study by Huang et al. [11].

Method	Dataset								
	Only classification			Classification & Segmentation					
	HIS [3]	ChestXray [20]	OCT17 [14]	BrainMRI [1, 2, 17]		LiverCT [4, 15]		RESC [9]	
				AC	AS	AC	AS	AC	AS
WinCLIP [13]	69.85	70.86	46.64	66.49	85.99	64.20	96.20	42.51	80.56
April-GAN [5]	72.36	57.49	92.61	76.43	**91.79**	70.57	97.05	75.67	85.23
MVFA-AD [11]	77.90	**71.11**	95.40	**78.63**	90.27	76.24	97.85	83.31	92.05
MLA-PL (Ours)	**79.89**	68.65	**96.35**	73.62	89.79	**76.77**	**98.82**	**84.50**	**92.27**

scales for evaluation: image-level for abnormality classification and pixel-level for abnormality segmentation.

Implementation Details. We employ the CLIP model using the ViT-L/14 architecture, processing input images at a resolution of 240. This model consists of 24 layers organized into 4 stages, each containing 6 layers. For training on Google Colab Pro, we utilize the Adam optimizer with a fixed learning rate of 1e-4 for visual adaptors and 1e-3 for the prompt learner, employing a batch size of 16 over 50 epochs.

4.2 Experimental Results

Zero-shot learning. For the zero-shot learning task, we will train on five datasets and evaluate on the remaining dataset. For instance, when evaluating on the ChestXray dataset, we will train on the HIS, OCT17, BrainMRI, LiverCT, and RESC datasets. Evaluation on the LiverCT, RESC, BrainMRI datasets includes both classification and segmentation tasks. Meanwhile, evaluation on ChestXray, HIS, and OCT17 datasets is limited to classification tasks only. The evaluation results of the model show that it surpasses the performance of the state-of-the-art model [11] on the HIS dataset (by more than 1.99%), OCT17 dataset (by more than 0.95%), LiverCT dataset (by 0.53% in AC and 0.97% in AS), and RESC dataset (by 1.19% in AC and 0.22% in AS). Detailed experimental results are provided in Table 1.

Few-shot learning. We conducted a comprehensive comparison of our proposed model against four previous methods: DRA [6], BGAD [22], April-GAN [5], and MVFA-AD [11]. In the Abnormality Classification (AC) task, our model consistently outperformed the others, especially in challenging datasets such as HIS and Chest X-ray, achieving notable improvements in accuracy. For the Abnormality Segmentation (AS) task, our model demonstrated superior performance in the Liver CT and RESC datasets, with segmentation results surpassing those of the compared models. Each model was evaluated on few-shot learning scenarios with k = 2, 4, 8, 16, demonstrating our model's robustness and effectiveness in low-data environments. The results clearly indicate that our model offers significant advancements in both detection and segmentation tasks, establishing it as a leading approach for medical image analysis. Detailed experimental results are provided in Table 2.

Table 2: In few-shot learning, when compared with state-of-the-art models in medical image anomaly detection, we observe that our model performs better in 16-shot. Additionally, in settings with shot = 2, 4, 8, 16, our model also achieves superior results compared to previous models on certain datasets. The results are evaluated using the AUC (Area Under the Curve) metric (%) for both anomaly classification (AC) and anomaly segmentation (AS). The performance of previous methods was collected from the study by Huang et al. [11].

Shots	Method	Dataset								
		Only classification			Classification & Segmentation					
		HIS [3]	ChestXray [20]	OCT17 [14]	BrainMRI [1, 2, 17]		LiverCT [4, 15]		RESC [9]	
					AC	AS	AC	AS	AC	AS
2-shot	DRA [6]	72.91	72.22	98.08	71.78	72.09	57.17	63.13	85.69	65.59
	BGAD [22]	-	-	-	78.70	92.42	72.27	**98.71**	83.58	92.10
	April-GAN [5]	69.57	69.84	**99.21**	78.45	94.02	57.80	95.87	89.44	96.39
	MVFA-AD [11]	**82.61**	81.32	97.98	92.72	96.55	81.08	96.57	91.36	**98.11**
	MLA-PL (Ours)	71.45	**86.07**	98.16	**92.73**	**96.98**	**81.81**	98.02	**92.17**	97.63
4-shot	DRA [6]	68.73	75.81	99.06	80.62	74.77	59.64	71.79	90.90	77.28
	BGAD [22]	-	-	-	83.56	92.68	72.48	98.88	86.22	93.84
	April-GAN [5]	76.11	77.43	99.41	89.18	94.67	53.05	96.24	94.70	97.98
	MVFA-AD [11]	82.71	81.95	99.38	**92.44**	**97.30**	81.18	**99.73**	**96.18**	98.97
	MLA-PL (Ours)	**83.05**	**82.13**	**99.66**	91.34	97.10	**84.50**	99.63	93.47	**99.02**
8-shot	DRA [6]	74.33	82.70	99.13	85.94	75.32	72.53	81.78	93.06	83.07
	BGAD [22]	-	-	-	88.01	94.32	74.60	99.00	89.96	96.06
	April-GAN [5]	81.70	73.69	**99.75**	88.41	95.50	62.38	97.56	91.36	97.36
	MVFA-AD [11]	85.10	83.89	99.64	**92.61**	**97.21**	85.90	**99.79**	96.57	**99.00**
	MLA-PL (Ours)	**87.38**	**84.12**	99.15	91.9	96.95	**89.39**	99.65	**97.14**	98.92
16-shot	DRA [6]	79.16	85.01	99.87	82.99	80.45	80.89	93.00	94.88	84.01
	BGAD [22]	-	-	-	88.05	95.29	78.79	99.25	91.29	97.07
	April-GAN [5]	81.16	78.62	**99.93**	94.03	96.17	82.94	99.64	95.96	98.47
	MVFA-AD [11]	82.62	85.72	99.66	94.40	97.70	83.85	**99.73**	97.25	99.07
	MLA-PL (Ours)	**85.56**	**87.39**	**99.93**	**94.51**	**97.91**	**92.75**	99.64	**97.37**	**99.34**

Finally, we visualize images from three datasets with segmentation data: BrainMRI [1, 2, 17], LiverCT [4, 15], and RESC [9]. The results are displayed in Figure 3. From left to right, there are three large columns, each containing three random images from the BrainMRI [1,2,17], LiverCT [4,15], and RESC [9] datasets. The first row shows the original medical images, while the next two rows display the heatmap and segmentation results for zero-shot and few-shot learning methods with k = 2, 4, 8, 16, the last row shows the images with ground truth.

In conclusion, our experimental results demonstrate the effectiveness and robustness of our proposed model in both zero-shot and few-shot learning scenarios for medical image analysis. In the zero-shot learning task, our model consistently outperforms the state-of-the-art methods on various datasets, showing significant improvements in both classification and segmentation tasks. Specifically, our model surpasses the performance of the leading model on the HIS, OCT17, LiverCT, and RESC datasets, establishing a new benchmark for these tasks. In the few-shot learning task, our model also shows superior performance compared to four established methods (DRA [6], BGAD [22], April-GAN [5], and MVFA-AD [11]) across six diverse datasets. The consistent improvements across different few-shot scenarios (k = 2, 4, 8, 16) highlight our model's ability to effectively learn from limited data. These results underscore the potential of

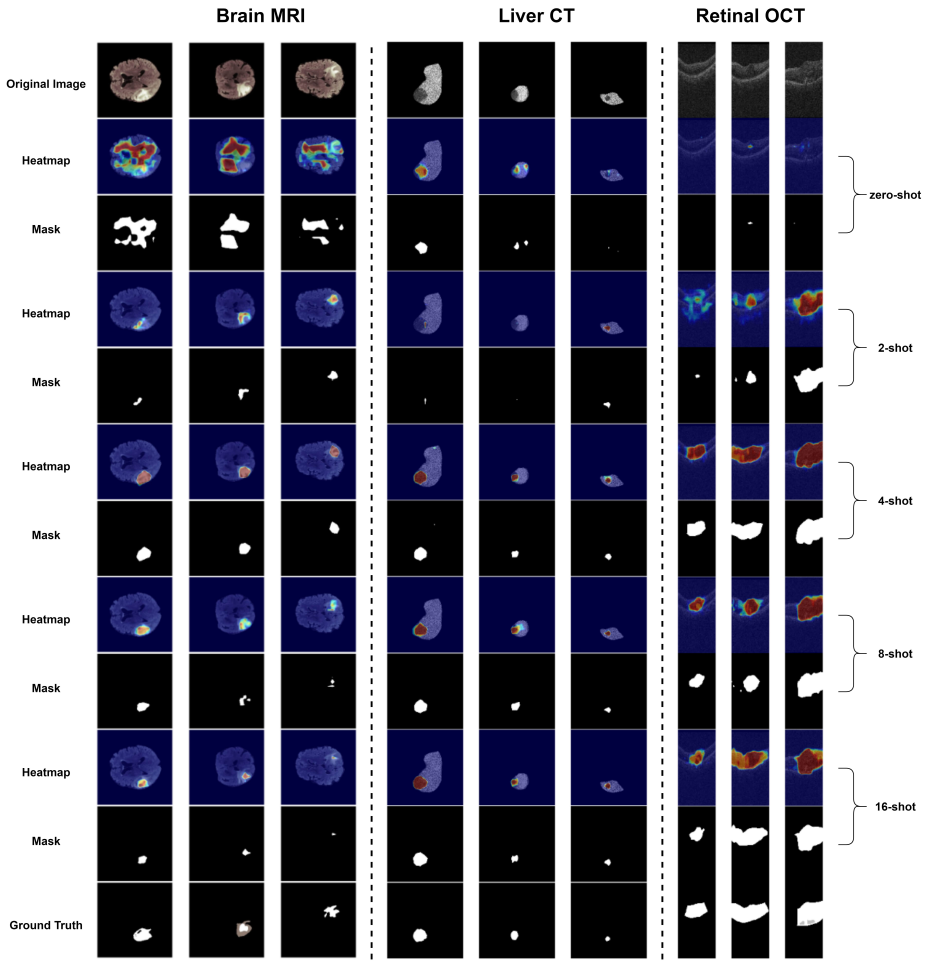

Fig. 3: We illustrate the segmentation results of the proposed model on three datasets: BrainMRI [1,2,17], LiverCT [4,15], and RESC [9]. The results are shown for both zero-shot and few-shot learning methods (with 2, 4, 8, and 16 shots).

our approach to significantly advance the field of medical image analysis, providing a robust and accurate solution for both abnormality classification and segmentation tasks.

5 Conclusion

In this study, we propose a method that combines Prompt Learning for CLIP's Text Encoder and Visual Adaptors for CLIP's Visual Encoder. We also demonstrate the strengths of CLIP for downstream tasks. Prompt Learning helps doctors bridge the gap between "natural" and "medical" semantics, enabling CLIP to grasp the meanings of medical images. Additionally, the Visual Adaptor aids CLIP in capturing high-level semantics for pixel-level segmentation. Our model has achieved promising results in both zero-shot and few-shot Abnormality Classification (AC) and Abnormality Segmentation (AS) tasks, indicating its potential for future research in this field.

In the future, we will experiment with additional methods based on the current idea and further improve certain parts. One promising direction we are exploring is the multi-level adapter forward prompt learner. Additionally, we are currently working on constructing a bone dataset for classification and segmentation tasks based on this model. However, this requires considerable time and support from medical professionals. We hope to publicly release a bone medical dataset in the future for the community, as this type of data is still relatively underrepresented in medical research.

Moreover, it would be valuable to investigate whether there are specific datasets on which this architecture might not perform well. Understanding such limitations could provide deeper insights into the strengths and weaknesses of the proposed method. We consider this an important direction for future research and aim to address it in subsequent studies.

Acknowledgments

This research is supported by research funding from University of Science, Vietnam National University - Ho Chi Minh City.

References

1. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021) 2, 8, 9, 10, 11
2. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017) 2, 8, 9, 10, 11

3. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., 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, 8, 9, 10
4. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023) 2, 8, 9, 10, 11
5. Chen, X., Han, Y., Zhang, J.: April-gan: 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) 2, 3, 8, 9, 10
6. Ding, C., Pang, G., Shen, C.: Catching both gray and black swans: Open-set supervised anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7388–7398 (2022) 8, 9, 10
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 5, 6
8. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024) 5
9. Hu, J., Chen, Y., Yi, Z.: Automated segmentation of macular edema in oct using deep neural networks. Medical image analysis **55**, 216–227 (2019) 2, 8, 9, 10, 11
10. Huang, C., Jiang, A., Feng, J., Zhang, Y., Wang, X., Wang, Y.: Adapting visual-language models for generalizable anomaly detection in medical images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 2, 3, 8
11. Huang, C., Jiang, A., Feng, J., Zhang, Y., Wang, X., Wang, Y.: Adapting visual-language models for generalizable anomaly detection in medical images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11375–11385 (2024) 8, 9, 10
12. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3942–3951 (2021) 2
13. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: Winclip: Zero-/few-shot anomaly classification and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19606–19616 (2023) 2, 3, 8, 9
14. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018) 2, 8, 9, 10
15. Koehler, G., Wald, T., Ulrich, C., Zimmerer, D., Jaeger, P.F., Franke, J.K., Kohl, S., Isensee, F., Maier-Hein, K.H.: Recyclenet: Latent feature recycling leads to iterative decision refinement. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 810–818 (2024) 2, 8, 9, 10, 11
16. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 8

17. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014) 2, 8, 9, 10, 11
18. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016) 8
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 2, 3, 4, 5, 6
20. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017) 2, 8, 9, 10
21. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022) 2
22. Yao, X., Li, R., Zhang, J., Sun, J., Zhang, C.: Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24490–24499 (2023) 8, 9, 10
23. Zhang, X., Xu, M., Qiu, D., Yan, R., Lang, N., Zhou, X.: Mediclip: Adapting clip for few-shot medical image anomaly detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 458–468. Springer (2024) 3
24. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022) 3
25. Zhou, Q., Pang, G., Tian, Y., He, S., Chen, J.: Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. arXiv preprint arXiv:2310.18961 (2023) 3