MadCLIP: Few-shot Medical Anomaly Detection with CLIP

Mahshid Shiri[1][0009−0008−5780−724X], Cigdem Beyan[1][0000−0002−9583−0087] ⋆, and Vittorio Murino[1,2][0000−0002−8645−2328]

[1] Department of Computer Science, University of Verona, Verona, Italy
[2] AI for Good (AIGO) Research Unit, Istituto Italiano di Tecnologia, Genoa, Italy
{mahshid.shiri, cigdem.beyan, vittorio.murino}@univr.it

Abstract. An innovative few-shot anomaly detection approach is presented, leveraging the pre-trained CLIP model for medical data, and adapting it for both image-level anomaly classification (AC) and pixel-level anomaly segmentation (AS). A dual-branch design is proposed to separately capture normal and abnormal features through learnable adapters in the CLIP vision encoder. To improve semantic alignment, learnable text prompts are employed to link visual features. Furthermore, SigLIP loss is applied to effectively handle the many-to-one relationship between images and unpaired text prompts, showcasing its adaptation in the medical field for the first time. Our approach is validated on multiple modalities, demonstrating superior performance over existing methods for AC and AS, in both same-dataset and cross-dataset evaluations. Unlike prior work, it does not rely on synthetic data or memory banks, and an ablation study confirms the contribution of each component. The code is available at https://github.com/mahshid1998/MadCLIP.

Keywords: Medical Anomaly Detection · CLIP · Adapters · Learnable prompts · Few-shot

1 Introduction

Medical anomaly detection (AD) involves identifying unusual patterns in medical data, a task complicated by the lack of a universal anomaly definition, inconsistent patterns, and noisy data from variably calibrated sensory devices. These challenges are magnified by the crucial role of AD in medical diagnosis, where high sensitivity is essential. Therefore, AD models in medicine must achieve exceptional performance for clinical reliability [8, 28, 34].

Overall, the AD task is approached from two main perspectives in the broader literature: (a) unsupervised techniques (e.g., [6, 26]) and (b) supervised methods (e.g., [13, 35, 32, 31]). Unsupervised methods detect anomalies by leveraging large datasets of normal samples, modeling the normal data distribution, and identifying anomalies as deviations. For example, PatchCore [25] compares test

⋆ Corresponding author

samples to a memory bank of normal embeddings and measures the nearest distance, while CFLOW-AD [9] models normal samples with a Gaussian distribution using normalizing flows.

While many methods rely on large datasets, real-world applications often include a few labeled anomalies, which provide valuable, application-specific insights and enable recent models to significantly improve the detection of similar anomalies [7]. In this context, supervised methods operate in a *few-shot AD* setting, where both normal and anomalous samples are limited, e.g., [7, 12, 27, 32]. However, the limited samples available during training, for both normal and anomalous classes, often fail to capture their full variability, restricting the model's ability to generalize to unseen cases [7].

Recently, CLIP [23] based methods have made significant strides in few-shot AD for medical images, e.g., [35, 13]. It is clear that, relying solely on CLIP [23] is insufficient, as its training focuses on aligning with the class semantics of foreground objects, limiting its ability to generalize and capture subtle visual abnormalities, and restricting its direct application in AD. Also, the substantial distribution shift between the data on which CLIP [23] was trained and medical images results in suboptimal performance when CLIP is applied directly to medical AD [13]. To effectively leverage CLIP for few-shot AD, it is crucial to address the domain gap and fine-tune or adapt CLIP [23] specifically for the medical AD task. For instance, MVFA [13] utilizes visual adapters in the form of fully connected layers, while MediCLIP [35] uses convolutional layers. On the other hand, several studies have shown that leveraging text modality as a representative of normal and abnormal classes can aid AD [16, 5, 35, 13]. Since anomaly descriptions might share similarities across different datasets, incorporating textual information reduces reliance only on visual data and, might enhance model performance, particularly in data-scarce scenarios such as few-shot learning. For instance, WinCLIP [16] uses a large set of artificial text prompts, while April-GAN [5] maps visual features extracted from CLIP [23] onto the linear space of text features in addition to using several memory banks. In medical AD, MVFA [13] builds on the principles of April-GAN [5] by employing multi-level adaptation of CLIP and utilizing fixed prompts. Unlike MVFA [13], WinCLIP [16], and April-GAN [5] using fixed prompts, MediCLIP [35] adopts the learnable prompts approach from [36]. Learnable prompts offer a key advantage over fixed prompts, which require careful design and expert knowledge for medical scenarios [35]. Besides visual feature adaptation and text prompts, another common approach is using memory banks (e.g., [5, 13]), though this strategy is relatively costly and often fails to generalize well. Some, e.g., [35], generate extensive synthetic data for the abnormal class to improve generalization.

Our approach, **MadCLIP** extends CLIP [23] with a two-branch architecture using adapters to capture normal and abnormal visual features. We further leverage the role of text in data-scarce scenarios, using it to represent both normal and abnormal distributions separately. MadCLIP is thus designed to learn these distributions from both visual data and text, aiming to obtain a clear distinction between normal and abnormal patterns. As a result, complementary

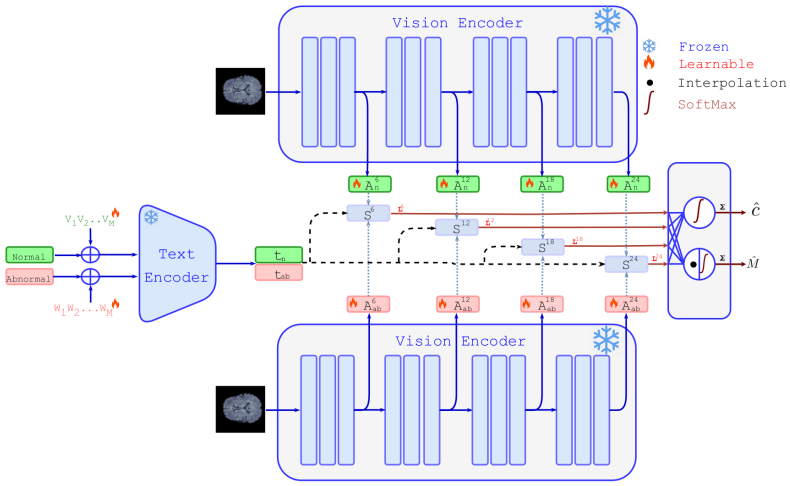

Fig. 1. Overview of MadCLIP: A dual-branch design integrates adapters A_n and A_{ab} into CLIP's vision encoder to separately capture normal and abnormal features. Learnable text prompts $V_1, \ldots, V_M$ and $W_1, \ldots, W_M$ encode complementary semantics for AD. The outputs are image-level AC $\hat{c}$ and AS mask $\hat{M}$.

branches exchange signals, and two sets of learnable prompts enable the model to capture distinctive descriptions for both normal and abnormal patterns. In detail, MadCLIP models normal and abnormal representations separately within a dual optimization process, maximizing multimodal (i.e., text and vision) similarity within each class while minimizing it between classes, thereby simplifying decision-making by subtracting (i.e., contrasting) learned feature representations to achieve better class separation. This enhances AD performance, particularly in cross-dataset settings, while also our pipeline avoids the need for memory banks or additional synthetic data. To perform image-text alignment, unlike the standard Softmax-based loss, we use SigLip [33], justified in the next section, where we also show its performance benefits with an ablation study. MadCLIP was evaluated on six datasets across five medical modalities using both a standard and cross-dataset approach. It outperforms state-of-the-art (SOTA) methods in anomaly classification (AC) and segmentation (AS).

The main contributions are: **(1)** A novel few-shot AD architecture with multi-level adapters, each focusing on either normal or abnormal instances, enhanced by a dual optimization objective utilizing learnable text embeddings for better separation. This approach does not require extensive synthetic data or memory banks, unlike SOTA methods. **(2)** This is the first application of SigLIP loss [33] in medical AD, proving its effectiveness. **(3)** Strong generalization and improved performance are demonstrated through extensive validation and cross-dataset evaluation across diverse medical modalities and anatomical areas.

2 Method

Few-shot medical AD is based on a dataset consisting of tuples $\{(x, c, M)\}$, where each $x \in \mathbb{R}^{h \times w \times 3}$ represents a training image with spatial dimensions $h \times w$, $c \in \{0, 1\}$ denotes the image-level AC label (1 for anomalous, 0 for normal), and, when available, $M \in \{0, 1\}^{h \times w}$ provides the pixel-level AS map. The training set is balanced, ensuring that $|\{x \mid c = 0\}| = |\{x \mid c = 1\}|$. Given a test image x_{test}, the model predicts both AC and AS. Building on this setup, we propose MadCLIP (see Fig. 1), a novel approach that leverages CLIP [23], incorporating multi-level visual features adaptation and learnable text prompts to enhance performance. MadCLIP is a dual-branch architecture that learns separate multi-modal representations for normal and abnormal samples. Below, we provide a detailed description of the components employed in MadCLIP.

Vision Adapters. MadCLIP employs adapters (denoted as A_n^i and A_{ab}^i for normal and abnormal samples, respectively) within the CLIP vision encoder, which is pre-trained on natural images [23], to effectively adapt it for medical imaging and the two target tasks: AC and AS. Using adapters follows the findings of [13], which shows that it is preferable to traditional fine-tuning, as it helps avoid overfitting due to high model complexity and limited data. Our adapters are integrated while keeping the backbone frozen. In detail, for an input image x, we extract the i-th layer feature from the CLIP vision encoder, denoted as $I^i(x) \in \mathbb{R}^{G \times d}$. Here, G represents the grid size, d is the feature dimension, and $i \in \{6, 12, 18, 24\}$. Learnable adapters consist of two transformation stages. The first stage focuses on addressing the domain gap between natural and medical images. Since CLIP embeddings are inherently optimized for object-level tasks in natural images, they may not directly align with the requirements of medical AD. To bridge this gap, we introduce a shared linear transformation layer (W_{shared}^i) that refines the CLIP embeddings, mapping them into a feature space better suited for medical AD, formulated as $F_{\text{shared}}^i(I^i(x)) = \text{ReLU}(W_{\text{shared}}^i I^i(x))$. The second stage, given the need to perform both AC and AS, focuses on extracting features specialized for each task, we apply two distinct linear transformation heads on top of the shared features: the first head, dedicated to AC, captures high-level characteristics essential for detecting anomalies, while the second head, tailored for AS, emphasizes fine-grained spatial details, expressed as $F_{\text{Det}}^i(I^i(x)) = \text{ReLU}\Big(W_{\text{det}}^i\, F_{\text{shared}}^i(I^i(x))\Big)$ and $F_{\text{Seg}}^i(I^i(x)) = \text{ReLU}\Big(W_{\text{Seg}}^i\, F_{\text{shared}}^i(I^i(x))\Big)$.

Learnable Prompts. Prior works have shown that well-designed text prompts in CLIP can encapsulate rich semantic information, resulting in more reliable and transferable representations for several tasks [20]. By using learnable prompts, e.g., [36], we can eliminate the complexity of manually engineered prompts, which we argue that it leads to better generalization of the resulting text embeddings to medical imaging tasks. Additionally, we posit that these prompts can provide complementary signals between our dual branches, enhancing AD

performance. Standard prompts such as "A photo of a [CLS]" primarily capture the overall semantic content of images, which often fails to reflect the subtle, domain-specific details found in medical imaging [37, 13]. To address this limitation, we develop a template for the normal (p_n) and abnormal (p_{ab}) classes as: $p_n = [V_1][V_2]\dots[V_M]$[CLS(normal)][Objective], $p_{ab} = [W_1][W_2]\dots[W_M]$ [CLS(abnormal)][Objective] where $[V_i]$ and $[W_i]$ denote learnable token embeddings, [CLS(normal)] and [CLS(abnormal)] are fixed class embeddings, and [Objective] encodes the fixed semantic context of the target modality (e.g., Brain). To further enhance AD, we leverage an ensemble of text prompts by incorporating multiple synonyms for *normal* (e.g. flawless, unblemished) and *abnormal* (e.g. with a flaw, disease). Each synonym generates a distinct prompt, leading to two prompt sets, $P_n = \{p_{n_1}, p_{n_2}, \dots, p_{n_k}\}$ and $P_{ab} = \{p_{ab_1}, p_{ab_2}, \dots, p_{ab_k}\}$, where k represents the number of synonyms. By aggregating the diverse textual prompts from these sets, we obtain two final prompts, t_n and t_{ab}.

Dual Branch Architecture. In few-shot AD, normal samples x_n are assumed to be drawn from an unknown distribution D_n (i.e., $x_n \sim D_n$), while anomalous samples x_{ab} originate from a distinct, typically unknown distribution D_{ab}. These two distributions are roughly complementary such that $D_{ab} = 1 - D_n$, based on the assumption that anomalies are defined as deviations from normal data. Our dual-branch architecture processes these distributions via two specialized branches.The so-called *normality branch* extracts features from normal samples and aligns them with a learnable text prompt t_n, capturing the behavior of D_n. In parallel, the so-called *abnormality branch* processes abnormal samples and aligns the extracted features with a complementary prompt t_{ab}. Each branch is trained to maximize the cosine similarity ($cosSIM$) between its visual features and the corresponding prompt while minimizing similarity with the opposing prompt (i.e., representative of the opposite class).

Formally, for a normal sample x_n, adapters A_n^i produce feature representations $O_n^i = A_n^i\big(I^i(x_n)\big)$. The dual optimization objective for normal samples is to maximize $\text{cosSIM}(O_n^i, t_n) - \text{cosSIM}(O_n^i, t_{ab})$, which, assuming cosine similarity approximates the dot product, simplifies to $\max\big[\mathbf{O_n^i \cdot t_n - O_n^i \cdot t_{ab}}\big]$. An analogous formulation is used for abnormal samples: $\max\big[\mathbf{O_{ab}^i \cdot t_{ab} - O_{ab}^i \cdot t_n}\big]$. These objectives aim to enforce a clear separation between the normal and abnormal feature spaces. Furthermore, at each feature layer i, the normality and abnormality scores are computed as $S_n^i = \big[O_n^i \cdot t_n - O_n^i \cdot t_{ab}\big]$ and $S_{ab}^i = \big[O_{ab}^i \cdot t_{ab} - O_{ab}^i \cdot t_n\big]$ and then concatenated into a single vector $S^i = \big[S_n^i, S_{ab}^i\big]$. During inference, as S is in patch-level, for AC we calculate the mean score over all patches, and for AS as we need to match the input size, we use bilinear interpolation to obtain image-level score vectors and aggregate them across layers, i.e., $\hat{M}^i = \text{SoftMax}\big(\text{Interpolate}(S^i)\big)$, $\hat{c}^i = \text{Mean}\big(\text{SoftMax}(S^i)\big)$ where $\hat{M} = \frac{1}{|i|}\sum_i \hat{M}^i$ is the predicted anomaly map and $\hat{c} = \frac{1}{|i|}\sum_i \hat{c}^i$ is predicted anomaly score. Here, $|i|$ refers to the total number of feature levels at which adapters are integrated into the visual encoder. This multi-layer adaptation aims to effectively integrate complementary information from both branches for ro-

bust AD.

Loss Function. The composite loss function we use at feature level i is $L^i = \lambda_1 \text{Dice}(\hat{M}^i, M) + \lambda_2 \text{Focal}(\hat{M}^i, M) + \lambda_3 \text{SigLip}(\hat{c}^i, c)$ where $\hat{M}^i$ represents the predicted anomaly map and $\hat{c}^i$ is the predicted anomaly score at the i'th feature level, M is the ground truth mask, and c denotes the image-level anomaly label. The hyperparameters λ_1, λ_2, and λ_3 are fixed to 1. $Dice(.,.)$, $Focal(.,.)$, and $SigLip(.,.)$ correspond to Dice [22], Focal [24], and a sigmoid-based loss for text-image alignment [33]. The Dice and Focal losses are necessary for AS, and particularly Focal loss is preferred as there is a significant class imbalance at the pixel level, with anomalous pixels being greatly outnumbered by normal pixels. Instead, for the AC task, given the balanced classes, the adaptation of SigLip [33] is sufficient. The overall loss L is computed as the sum of losses across all feature levels: $L = \sum_i L^i$.

We use **SigLip loss** [33] instead of the original CLIP loss[23] because our architecture introduces two learnable text embeddings, each linked to multiple images. Specifically, image similarity is computed across both normal and abnormal prompts, while each prompt is compared against all images. This results in a similarity matrix capturing multiple valid associations rather than a strict diagonal mapping. While CLIP loss can compute image similarity across all texts, it cannot directly handle text similarity across multiple images, as each text embedding corresponds to multiple images in a batch. In contrast, SigLip loss processes image-text pairs independently, naturally supporting our many-to-one setup. This distinguishes our approach from previous work in medical AD [31, 11, 35, 13], with its contribution experimentally validated below.

3 Experimental Analysis and Results

We follow the latest SOTA: MVFA [13], using a medical AD benchmark that spans five modalities and six datasets: brain MRI [1, 2, 21], liver CT [4, 19], retinal OCT (composed of two datasets; OCT17 [17], and RESC [10]), chest X-ray (Chest) [30], and digital histopathology (HIS) [3]. BrainMRI, LiverCT, and RESC are used for both AC and AS, while OCT17, Chest, and HIS are relevant only for AC. We use the area under the Receiver Operating Characteristic curve (AUC), the standard medical AD metric, to report AUC for AC and AUC for AS.

Implementation Details. We use the CLIP model with the ViT-L/14 architecture and 240-pixel input images, as in [13]. The model has 24 layers and the adapters were applied to the 6th, 12th, 18th, and 24th layers. Training is performed with the Adam optimizer with the learning rate of $1e^{-3}$, batch size 16, for 60 epochs. Augmentation follows the strategy outlined in [13].

Comparisons with SOTA. Table 1 presents the results of MadCLIP alongside SOTA. Methods labeled as unsupervised (referred to as Unsup) rely on large auxiliary datasets containing only normal samples, while *few-shot* methods utilize a

Table 1. Comparisons with SOTA in terms of AUC (%). Few-shot models use 16 samples per class. Best results are **bold**, second-best underlined.

	Method	Source	HIS	Chest	OCT17	BrainMRI		LiverCT		RESC		Average	
			AC	AC	AC	AC	AS	AC	AS	AC	AS	AC	AS
Unsup	CFLOWAD [9]	WACV22	54.54	71.44	85.43	73.97	93.52	49.93	92.78	74.43	93.75	68.29	93.35
	RD4AD [6]	CVPR22	66.59	67.53	97.24	89.38	96.54	60.02	95.86	87.53	96.17	78.04	96.19
	PatchCore [25]	CVPR22	69.34	75.17	98.56	91.55	96.97	60.40	96.58	91.50	96.39	81.09	96.65
	MKD [26]	CVPR22	77.74	81.99	96.62	81.38	89.54	60.39	96.14	88.97	86.60	81.18	90.76
Few shot	DRA [7]	CVPR22	79.16	85.01	99.87	82.99	80.45	80.89	93.00	94.88	84.01	87.13	85.82
	BGAD [32]	CVPR23	-	-	-	88.05	95.29	78.79	99.25	91.29	97.07	-	97.20
	APRIL-GAN [5]	CVPRw23	81.16	78.62	**99.93**	94.03	96.17	82.94	99.64	95.96	98.47	88.77	98.09
	MediCLIP [35]	MICCAI24	70.22	69.74	96.37	91.56	98.08	79.31	98.95	86.51	94.07	82.28	97.03
	MVFA [13]	CVPR24	82.62	85.72	99.66	94.40	97.70	83.85	99.73	97.25	99.07	90.58	98.83
	MadCLIP (Ours)		**90.14**	**88.15**	99.71	**95.9**	**97.97**	**91.46**	**99.74**	**99.11**	**99.45**	**94.08**	**99.05**

fixed set of 16 normal and abnormal samples. MadCLIP outperforms all SOTA across multiple datasets, except for OCT17 [17], where April-GAN [5] achieves better results. The slightly better performance of April-GAN [5] on OCT17 [17] is likely due to the low inter-sample variability of the dataset, which is known to benefit memory bank-based methods that operate by comparing test samples with stored representations from the training set. On average, MadCLIP achieves best overall performance, surpassing the second-best method by 3.5% in AC and 0.22% in AS. The performance gain of MadCLIP can reach up to 25.79% in AC and 13.23% in AS. Table 2 further compares few-shot methods for different numbers of normal/abnormal samples. On average, independent of the number of samples, MadCLIP performs the best, except for the OCT17 [17], where all methods perform very similarly. Overall, as expected, the performance of the methods increases as more samples are added to the training data. Still, in the extreme case where only 2 normal samples and 2 abnormal samples are available, MadCLIP outperforms the others in 7 out of 9 cases.

Ablation studies. We conducted ablation studies on both the AC and AS tasks, reporting average results over three different seeds and six datasets to assess the overall effectiveness of MadCLIP. **(a)** The impact of prompt design was evaluated by replacing our learnable prompt tokens (i.e., $[V_1][V_2]\dots[V_M]$ and $[W_1][W_2]\dots[W_M]$) with the hand-crafted templates used in [13, 16]. This substitution led to performance drops of 1.28% for AC and 1.3% for AS, highlighting the contribution of our learnable prompts. **(b)** Substituting the `[Objective]` term with "medical image" resulted in drops of 1.71% in AC and 1.01% in AS, further demonstrating the benefits of learnable prompts and context-specific information. We further examined our dual-branch design by performing two ablations. **(c)** removing one set of adapters (i.e., leaving 4 shared adapters for both normal and anomaly classes, instead of the 8 in MadCLIP) resulted in declines of 1.14% for AC and 1.06% for AS. **(d)** eliminating the signal from the opposite class while calculating S_n^i and S_{ab}^i (i.e., excluding the subtraction term from the calculation of the mentioned formulas), led to decreases of 1.41% for AC and 1.24% for AS. **(e)** We replaced SigLip Loss with the CLIP-based SoftMax loss, resulting in a performance drop of 1.48% for AC and 1.1% for AS.

Table 2. Comparisons with few-shot SOTA for 2, 4, and 8 shots per class (AUC %). Results for 16 shots are in Table 1. Best results are **bold**, second-best underlined.

	Method	Source	HIS	Chest	OCT17	BrainMRI		LiverCT		RESC		Average	
			AC	AC	AC	AC	AS	AC	AS	AC	AS	AC	AS
2	DRA [7]	CVPR22	72.91	72.22	98.08	71.78	72.09	57.17	63.13	85.69	65.59	76.3	66.93
	BGAD [32]	CVPR23	-	-	-	78.70	92.42	72.27	98.71	83.58	92.10	-	94.41
	APRIL-GAN [5]	CVPRw23	69.57	69.84	**99.21**	78.45	94.02	57.80	95.87	89.44	96.39	77.38	95.42
	MediCLIP [35]	MICCAI24	64.49	61.69	93.4	85.13	97.39	68.48	97.09	83.96	96.01	76.2	96.83
	MVFA [13]	CVPR24	82.61	81.32	97.98	92.72	96.55	81.08	96.57	91.36	**98.11**	87.84	97.07
	MadCLIP (Ours)		**83.62**	**84.56**	99.06	**93.93**	**97.92**	**84.48**	**99.39**	**95.09**	97.18	**90.12**	**98.16**
4	DRA [7]	CVPR22	68.73	75.81	99.06	80.62	74.77	59.64	71.79	90.90	77.28	79.12	74.61
	BGAD [32]	CVPR23	-	-	-	83.56	92.68	72.48	98.88	86.22	93.84	-	95.13
	APRIL-GAN [5]	CVPRw23	76.11	77.43	**99.41**	89.18	94.67	53.05	96.24	94.70	97.98	81.64	96.29
	MediCLIP [35]	MICCAI24	70.85	56.83	89.07	83.82	96.86	81.53	98.61	87.52	96.65	78.27	97.37
	MVFA [13]	CVPR24	**82.71**	81.95	99.38	92.44	97.30	81.18	**99.73**	96.18	**98.97**	88.97	98.66
	MadCLIP (Ours)		80.05	**88.10**	99.37	**95.25**	**97.90**	**82.97**	99.29	**96.62**	98.9	**90.39**	**98.69**
8	DRA [7]	CVPR22	74.33	82.70	99.13	85.94	75.32	72.53	81.78	93.06	83.07	84.61	80.05
	BGAD [32]	CVPR23	-	-	-	88.01	94.32	74.60	99.00	89.96	96.06	-	96.46
	APRIL-GAN [5]	CVPRw23	81.70	73.69	**99.75**	88.41	95.50	62.38	97.56	91.36	97.36	82.88	96.80
	MediCLIP [35]	MICCAI24	69.8	72.08	95.69	92.29	**98.02**	86.32	98.32	88.82	95.98	84.17	97.44
	MVFA [13]	CVPR24	85.10	83.89	99.64	92.61	97.21	85.90	99.79	96.57	**99.00**	90.61	98.66
	MadCLIP (Ours)		**87.45**	**83.90**	99.14	**95.17**	**98.02**	**89.31**	**99.81**	**97.16**	98.85	**92.02**	**98.89**

Table 3. Cross-dataset evaluation. AC is reported for all datasets. Best are **bold**.

Source	Chest		BrainMRI	OCT17	RESC	AVG
Target	NIHChest [29]	CheXpert [14]	ADNI [15]	OCTDL [18]		
MVFA [13]	61.91	80.41	53.94	88.47	87.94	74.53
MadCLIP (Ours)	**62.44**	**81.99**	**57.35**	**90.74**	**88.09**	**76.12**

(f) The use of a single common head instead of separate heads, F_{Det} and F_{Seg}, for the AC and AS tasks, respectively resulted in performance drops of 1.43% for AC and 3.41% for AS. To sum up, these ablation studies validate our design choices, demonstrating their positive contribution to both AC and AS tasks.

Cross-dataset analysis. MadCLIP is compared with the best counterpart: MVFA [13] to assess generalization in a cross-dataset setting. Each model was trained on 16 samples from the datasets described above and tested on unseen target datasets of the same modality, as listed in Table 3. As seen, MadCLIP consistently outperforms MVFA across all individual datasets, demonstrating superior cross-dataset generalization. Notably, MadCLIP achieves an average performance of 76.12% compared to 74.53% for MVFA.

4 Conclusion

We introduced a few-shot AD architecture that leverages CLIP with multi-level adapters and prompt learning to model normal and abnormal classes separately. Our dual-objective strategy, formulated through subtraction, incorporates a contrastive effect by encouraging similarity within the same class and dissimilarity

between opposing class. By integrating SigLIP loss, we further refine this separation process, as it can handle many-to-one relationship of images and learnable unpaired texts. Extensive validation across diverse datasets demonstrates superior performance and strong generalization over SOTA methods, underscoring our approach's robustness for medical AD. However, a current limitation of MadCLIP is the assumption that learnable adapters and prompts are sufficient to bridge the gap between medical images and textual descriptions, while the modality gap remains unaddressed explicitly. Future work will explore the method's zero-shot potential and extend it to multi-modal AD, handling diverse training modalities and unseen anatomical regions simultaneously.

Acknowledgments. We acknowledge the financial support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baid, U., Ghodasara, S., Mohan, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv:2107.02314 (2021)
2. Bakas, S., Akbari, H., Sotiras, A., et al.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017)
3. Bejnordi, B.E., Veta, M., Van Diest, P.J., 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)
4. Bilic, P., Christ, P., Li, H.B., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023)
5. Chen, X., Han, Y., Zhang, J.: A zero-/fewshot 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:2305.17382 **2**(4) (2023)
6. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: CVPR. pp. 9737–9746 (2022)
7. Ding, C., Pang, G., Shen, C.: Catching both gray and black swans: Open-set supervised anomaly detection. In: CVPR. pp. 7388–7398 (2022)
8. Fernando, T., Gammulle, H., Denman, S., Sridharan, S., Fookes, C.: Deep learning for medical anomaly detection–a survey. ACM Comp. Surveys **54**(7), 1–37 (2021)
9. Gudovskiy, D., Ishizaka, S., Kozuka, K.: Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In: WACV. pp. 98–107 (2022)
10. Hu, J., Chen, Y., Yi, Z.: Automated segmentation of macular edema in oct using deep neural networks. Medical image analysis **55**, 216–227 (2019)
11. Hua, L., Luo, Y., Qi, Q., Long, J.: Medicalclip: Anomaly-detection domain generalization with asymmetric constraints. Biomolecules **14**(5), 590 (2024)

12. Huang, C., Guan, H., Jiang, A., Zhang, Y., Spratling, M., Wang, Y.F.: Registration based few-shot anomaly detection. In: ECCV. pp. 303–319. Springer (2022)
13. Huang, C., Jiang, A., Feng, J., Zhang, Y., Wang, X., Wang, Y.: Adapting visual-language models for generalizable anomaly detection in medical images. In: CVPR. pp. 11375–11385 (2024)
14. Irvin, J., Rajpurkar, P., Ko, M., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI. pp. 590–597 (2019)
15. Jack Jr, C.R., Bernstein, M.A., Fox, N.C., et al.: The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging **27**(4), 685–691 (2008)
16. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: Winclip: Zero-/few-shot anomaly classification and segmentation. In: CVPR (2023)
17. Kermany, D.S., Goldbaum, M., Cai, W., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018)
18. Kulyabin, M., Zhdanov, A., Nikiforova, A., et al.: Octdl: Optical coherence tomography dataset for image-based deep learning methods. Scientific data **11** (2024)
19. Landman, B., Xu, Z., Igelsias, J., et al.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5, p. 12 (2015)
20. Li, Y., Goodge, A., Liu, F., Foo, C.S.: Promptad: Zero-shot anomaly detection using text prompts. In: WACV. pp. 1093–1102 (2024)
21. 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)
22. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: fourth international Conf. on 3D vision. pp. 565–571. Ieee (2016)
23. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
24. Ross, T.Y., Dollár, G.: Focal loss for dense object detection. In: CVPR. pp. 2980–2988 (2017)
25. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: CVPR. pp. 14318–14328 (2022)
26. Salehi, M., Sadjadi, N., Baselizadeh, S., et al.: Multiresolution knowledge distillation for anomaly detection. In: CVPR. pp. 14902–14912 (2021)
27. Sheynin, S., Benaim, S., Wolf, L.: A hierarchical transformation-discriminating generative model for few shot anomaly detection. In: ICCV. pp. 8495–8504 (2021)
28. Su, J., Shen, H., Peng, L., Hu, D.: Few-shot domain-adaptive anomaly detection for cross-site brain images. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(3), 1819–1835 (2021)
29. Summers, R.: Nih chest x-ray dataset of 14 common thorax disease categories. NIH Clinical Center: Bethesda, MD, USA (2019)
30. 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: CVPR. pp. 2097–2106 (2017)
31. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv:2210.10163 (2022)

32. Yao, X., Li, R., Zhang, J., Sun, J., Zhang, C.: Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In: CVPR. pp. 24490–24499 (2023)
33. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV. pp. 11975–11986 (2023)
34. Zhang, J., Xie, Y., Pang, G., et al.: Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection. IEEE transactions on medical imaging **40**(3), 879–890 (2020)
35. Zhang, X., Xu, M., Qiu, D., Yan, R., Lang, N., Zhou, X.: Mediclip: Adapting clip for few-shot medical image anomaly detection. In: MICCAI. pp. 458–468. Springer (2024)
36. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Comp. Vision (2022)
37. Zhou, Q., Pang, G., Tian, Y., He, S., Chen, J.: Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. arXiv:2310.18961 (2023)