2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

FusionCLIP-AD: Hierarchical Global-Local Adaptation with Learnable Embeddings for Robust Medical Image Anomaly Detection

Hongwei Li[1,2] and S. Kevin Zhou[1,2,3*]

[1] School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China (USTC)，Hefei Anhui, 230026, China
[2] Center for Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE), Suzhou Institute for Advance Research, USTC, Suzhou Jiangsu, 215123, China
[3] State Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, China

Abstract. Although CLIP-based few-shot learning has shown promise in anomaly detection, it still exhibits notable limitations in medical imaging applications: fixed prompt mechanisms are difficult to finely adapt to domain differences, and the lack of collaborative modeling between local and global features results in loss of holistic information. This paper proposes a novel hierarchical adaptation framework: 1) Integration of global and local features to effectively capture potential details and comprehensive information in medical images, and 2) Multi-level learnable anomaly prompts dynamically constructed in the embedding space. By learning fused features and prompts across different layers, the model flexibly and accurately addresses complex scenarios in medical imaging. Experimental results demonstrate that the proposed method significantly enhances CLIP's few-shot learning performance in medical image anomaly detection tasks. Our method achieves state-of-the-art performance on LiverCT with **85.55% AUROC** under 4-shot settings, surpassing prior arts like MVFA (81.18%).

Keywords: Vision-Language Model, Few-Shot Anomaly Detection, Medical Image Analysis.

1 Introduction

In recent years, with the rapid development of deep learning and large-scale pre-trained models, vision-language models [1, 14, 18] (e.g., CLIP [18]) have shown significant potential in multimodal learning tasks. Particularly in few-shot learning, CLIP enhances model performance under limited annotation through cross-modal learning capabilities. Few-shot learning [8, 11, 19, 23] not only reduces dependence on large-scale labeled datasets but also provides new solutions for data-scarce domains like medical imaging, where annotation is expensive and time-consuming.

*Corresponding author

However, despite CLIP's success in general domains, its adaptability and performance in medical imaging remain limited. The differences in visual features and structures between medical and natural images lead to suboptimal CLIP performance in medical anomaly detection. Existing CLIP-based few-shot methods rely on fixed prompts for feature extraction and inference, but such rigid prompts lack flexibility to fully capture medical image characteristics. Existing approaches in few-shot scenarios still suffer from excessive reliance on local information and loss of global context, thereby compromising model robustness and accuracy.

To address these issues, this paper proposes the FusionCLIP-AD framework. By fusing class tokens and patch tokens, the model integrates global and local features, enhancing adaptability to medical imaging. Additionally, multi-level learnable anomaly prompts are designed to dynamically align with hierarchical visual features, improving detection performance.

2 Related work

Vision-Language Models (VLMs), which learn aligned image-text representations through joint training, offer novel paradigms for few-shot medical diagnosis. Early works such as ConVIRT [24] demonstrated the effectiveness of cross-modal learning in chest X-ray classification by pretraining models on radiology reports. CLIP, trained via contrastive learning on 400 million web-crawled image-text pairs, exhibits strong zero-shot transfer capabilities. However, its direct application to medical imaging is limited by domain gaps between natural and medical images (e.g., grayscale monotony in X-rays, multi-modality characteristics in MRI).

WinCLIP [13] is a zero-shot anomaly detection method based on the CLIP model. It extracts multi-scale window features via a sliding window strategy and calculates similarity between these windows and anomaly prompts. While WinCLIP's key strength lies in its zero-shot learning capability (requiring no labeled data), its computational efficiency is low due to repeated calculations across window scales during inference. To improve performance, WinCLIP+ introduces few-shot learning with a memory bank for reference similarity computation. Although this enhances performance, sliding window operations of references incur higher computational costs. APRIL-GAN [7] combines handcrafted prompt templates and learnable linear layers to adapt CLIP's patch tokens for anomaly detection. It achieved 4th place in the few-shot track and first place in the zero-shot track at the CVPR 2023 challenge. The success of APRIL-GAN demonstrates that few-shot CLIP-based anomaly detection methods necessitate a few number of learnable parameters to maintain effectiveness. MVFA [12] proposes a multi-layer adapter to adapt CLIP from natural to medical images. By aligning visual features, it enhances CLIP's medical imaging adaptability and improves anomaly detection performance. However, APRIL-GAN and MVFA rely solely on patch tokens, lacking holistic image-level information critical for anomaly detection. While WinCLIP utilizes both class tokens and patch tokens, its class tokens (typically aggregated from 4–9 patches) fail to capture comprehensive image-level semantics, weakening anomaly classification. A common limitation across these methods is their

reliance on fixed handcrafted prompts, which limits fine-grained anomaly detection capabilities.

3 Methodology

This chapter introduces the principles of FusionCLIP-AD, including the dual-path adapter, learnable anomaly prompts, as well as the training and inference procedures.

3.1 Global-Local Token Fusion

To fully leverage the output features of the vision encoder, we designed a dual-path adapter mechanism to better integrate and utilize both global and local information of images. Specifically, we selected distinct outputs from the CLIP vision encoder as inputs for the adapters. Our dual-path adapter framework comprises two types: the first type integrates the class token with patch tokens, while the second type exclusively preserves patch tokens.

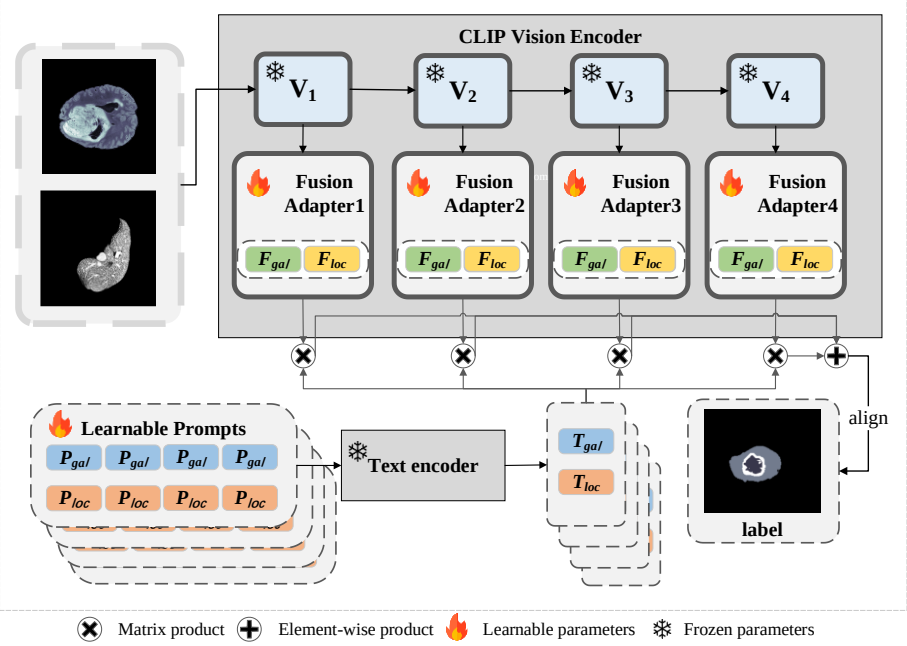

Fig. 1. The framework of FusionCLIP-AD

As illustrated in Figure 1, for an input image X, processing through different layers(denoted as V_1 to V_4) of the CLIP vision transformer [9, 20] (ViT) encoder yields multiple tokens, including N patch tokens and one class token.

$$\{token^{(l)}\}_{l=1}^{L} = \text{CLIP-ViT(X)}, \quad token^{(l)} \in R^{(N+1)\times d} \tag{1}$$

The patch tokens encapsulate localized details, whereas the class token represents holistic semantics. Our key innovation is merging each patch token and class token to create N fused tokens (Equation 1, where *l* indicates the layer index in ViT).

$$token_{fuse,i}^{(l)} = \text{Repeat}\left(token_{cls}^{(l)}, N\right) \parallel token_{i}^{(l)}, \quad \forall i \in [1, N] \tag{2}$$

In Equation 2, the newly generated token$^{(l)}_{fuse,i}$ integrate both global and local information. To prevent performance degradation caused by potential loss of local information in the fused tokens (which incorporate class tokens), we employ dual adapters for the original patch tokens and fused tokens, the first adapter:

$$F_{gal}^{(l)} = \text{ReLU}\left(\text{W}_{gal}^{(l)} \cdot token_{fuse}^{(l)}\right) \tag{3}$$

And patch token's adapter:

$$F_{loc}^{(l)} = \text{ReLU}\left(\text{W}_{loc}^{(l)} \cdot token_{\text{pat}}^{(l)}\right) \tag{4}$$

In Equation 3 and Equation 4, *F* refers to the features after passing through a linear layer. F_{gal} and F_{loc} represent fusion feature and local features, respectively. W_{gal} and W_{loc} represent linear projection, and "·" represents martrix product. The features obtained through Equation 3 and 4 share the same dimensionality as the text features. These aligned features are used to calculate similarity with the text features for anomaly detection.

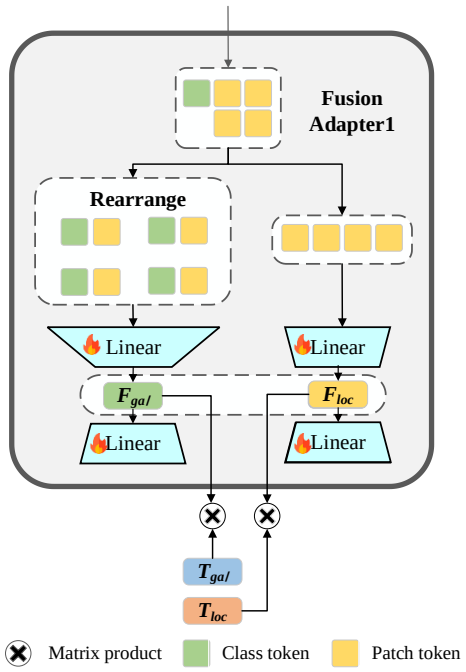

Fig. 2. The architectural of fusion adapter.

As figure 2 shows, we present a detailed architecture of the fusion adapter. Specifically, the Transformer Block outputs patch tokens and a class token. The class token is broadcast to all patch tokens, and both sets of tokens are then fed into separate adapter branches, yielding visual features F_{gal} and F_{loc}. These visual features are subsequently used to compute cosine similarity with the fusion-enhanced text embeddings T_{gal} and T_{loc}, respectively, which drive the final anomaly feature map generation.

3.2 Multi-level learnable anomaly prompts

Conventional anomaly detection methods rely on fixed manual templates designed for natural or industrial images. However, such fixed templates struggle to bridge the domain gap between natural and medical images, lack flexibility, and may cause visual feature adapters to overfit. To address these limitations, we propose multi-level learnable anomaly prompts.

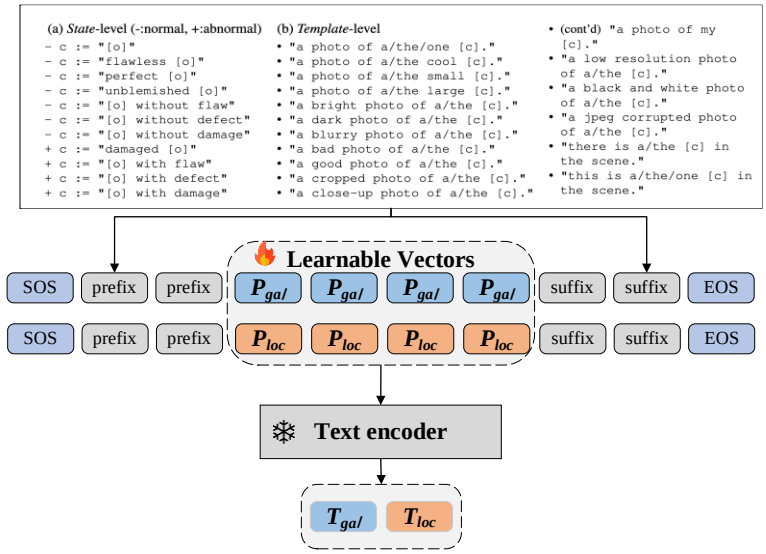

Fig. 3. The generation process of multi-level learnable anomaly prompts.

As shown in figure 3, Our framework first generates normal prompts and anomaly prompts using predefined templates (inspired by APRIL-GAN). These template-based prompts are processed through CLIP's tokenizer to obtain token embeddings. Next, we divide these token embeddings into two parts: the tokens before the object (such as "LiverCT") are treated as a non-learnable prefix, while the object token and subsequent tokens form a non-learnable suffix. It is important to note that both the prefix and suffix are fixed, and only the intermediate segment between them contains learnable vectors. These learnable vectors are categorized into P_{gal} and P_{loc}. By passing them through CLIP text encoder, we obtain encoded text embeddings T_{gal} and T_{loc}, which are aggregated from multiple vectors. We then compute cosine similarity between T_{gal}/T_{loc} and the visual features F_{gal}/F_{loc}. Additionally, distinct T_{gal} and T_{loc} are generated for each

layer's F_{gal} and F_{loc} to align with the varying semantic levels of visual features across different layers.

3.3 Train and Inference

After obtaining the **dual-path visual features (*F*)** and **learnable text features (*T*)**, we calculate the loss function to achieve cross-modal feature realignment. Our framework supports two types of anomaly detection tasks:

1 **Image-level detection**: Determining whether an image contains lesions.
2 **Pixel-level segmentation**: Precisely localizing lesion regions.

For datasets with **only image-level labels**, we use the **Binary Cross-Entropy (BCE) loss**.

$$\mathcal{L}\text{bce} = \sum_{i\in\{gal,loc\}} \alpha_i \, \text{BCE}\left(\sigma\left(\frac{1}{|\mathcal{R}_i|}\sum_{(h,w)\in\mathcal{R}_i} F_i^{(h,w)}, T_i\right), label\right) \tag{5}$$

In Equation 5, we compute the dot products between each of the *h*x*w* F_{gal} and F_{loc} features and their corresponding *T*, then aggregate the results. The aggregated value is subsequently used to calculate the binary cross-entropy (BCE) loss against the ground-truth labels.For datasets with **pixel-level annotations**, we employ a hybrid loss combining **Focal Loss** (to address class imbalance) and **Dice Loss** (to optimize segmentation boundary consistency).

$$\mathcal{L}seg = \text{Dice}(\text{bilinear}(F,T), label) + \text{Focal}(\text{bilinear}(F,T), label) \tag{6}$$

It is noteworthy that the formulas mentioned above utilize only single-layer text and visual features. During actual training and inference, we employ multi-layer features for joint computation. In the inference stage, we adopt a dual-branch framework: a zero-shot branch and a few-shot branch, following the same configuration as APRIL-GAN (thus not reiterated here). Unlike WinCLIP, we do not directly use class tokens for anomaly detection. One key reason is that WinCLIP's sliding window mechanism involves overlapping value aggregation across multiple windows, an operation challenging to implement efficiently on GPUs and highly time-consuming. Both APRIL-GAN and MVFA exclusively utilize patch tokens to train anomaly detection models. In our experiments, we observed that models trained solely on class tokens underperform, while patch tokens demonstrate superior trainability—an indisputable fact. However, relying solely on patch tokens sacrifices global contextual information, which is critical for image-level anomaly assessment. Additionally, our learnable vectors in the embedding space contribute to finer-grained decision boundaries.

4 Experiment

This chapter presents our experimental setup, comparative results with state-of-the-art methods, ablation studies, and visualization of detection outcomes.

4.1 Experimental Setups

Datasets.
BMAD [4] (Benchmarks for Medical Anomaly Detection) is a specialized benchmark dataset constructed for medical image anomaly detection tasks. BMAD integrates six reorganized datasets from five distinct medical domains: MRI [2, 3, 17], CT [6, 16], OCT [10, 15], Chest X-ray [21, and histopathology [5]. Among these, the BrainMRI, LiverCT, and RESC datasets are utilized for both anomaly classification (AC) and anomaly segmentation (AS), while OCT17, Chest X-ray, and HIS are exclusively designed for AC tasks. This benchmark aims to standardize evaluation protocols and bridge the domain gap in medical anomaly detection research.

Competing Methods and Baselines.
In this study, we compare our method with CLIP, WinCLIP,APRIL-GAN,MVFA,and MedCLIP [22] under unified evaluation protocols for anomaly classification (AC) and anomaly segmentation (AS). 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. 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. It should be noted that to achieve optimal performance for each dataset, we employed dataset-specific learning rates during training.

4.2 Comparison with State-of-the-art Methods

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

method	source	HIS	Chest	OCT	Brain	Liver	RESC
CLIP	Open-CLIP	63.48	70.74	98.59	74.31	56.74	84.54
MedCLIP	EMNLP 2022	75.89	**84.06**	81.39	76.87	60.65	66.58
WinCLIP	CVPR 2023	67.49	70.00	97.89	66.85	67.19	88.83
APRIL-GAN	CVPR-VAND	76.11	77.43	**99.41**	89.18	53.05	94.70
MVFA	CVPR 2024	<u>82.71</u>	81.95	<u>99.38</u>	<u>92.44</u>	<u>81.18</u>	**96.18**
Ours	Ours	**83.48**	<u>82.07</u>	99.29	**92.77**	**85.55**	<u>95.80</u>

Our method demonstrates superior anomaly detection performance across all six medical imaging datasets in the BMAD benchmark. As shown in the table 1, our approach

achieves state-of-the-art results on four tasks: HIS (83.48%), ChestXray (82.07%), BrainMRI (92.77%), and LiverCT (85.55%), with a significant improvement of 7.59% over the suboptimal MedCLIP method (75.89%) on HIS and 4.37% on LiverCT. Notably, the original CLIP model pretrained on natural images exhibits substantial limitations in medical domains, as evidenced by its 56.74% AUROC on LiverCT, while our dual-path feature adaptation mechanism achieves 85.55% detection accuracy, validating the effectiveness of cross-domain feature realignment. All methods attain high performance on the OCT17 retinal dataset (Open-CLIP: 98.59%, Ours: 99.29%), indicating relatively well-defined anomaly patterns in this task. The minor performance gap on the RESC dataset (95.80% vs. MedCLIP's 96.18%) may stem from the morphological specificity of fundus lesions, which presents an optimization direction for future research. The experimental results comprehensively demonstrate that our proposed learnable embedding strategy and hierarchical feature fusion mechanism effectively bridge the domain gap between medical imaging and natural images.

Table 2. Comparisons with state-of-the-art few-shot anomaly detection methods with K=4. The AUCs (in %) for anomaly segmentation (AS) are reported. The best result is in bold, and the second-best result is underlined.

method	source	Brain	Liver	RESC
CLIP	Open-CLIP	93.44	97.20	95.03
MedCLIP	EMNLP 2022	90.91	94.45	88.98
WinCLIP	CVPR 2023	94.16	96.75	96.68
APRIL-GAN	CVPR-VAND	94.67	96.24	<u>97.98</u>
MVFA	CVPR 2024	**97.30**	<u>99.73</u>	**98.97**
Ours	Ours	<u>96.73</u>	**99.80**	97.68

As shown in the table 2, our approach achieves **99.80% AUROC** on the Liver dataset, surpassing the state-of-the-art method MVFA (CVPR 2024, 99.73%), which validates its superiority in liver CT anomaly detection. On the Brain and RESC datasets, our method attains competitive results of **96.73%** and **97.68%**, respectively, closely aligning with MVFA (Brain: 97.30%, RESC: 98.97%) and APRIL-GAN (RESC: 97.98%). Notably, the original CLIP model (OpenCLIP) performs poorly on RESC (95.03%), while our method significantly improves to 97.68% through cross-modal feature alignment. Although MVFA achieves the highest accuracy on RESC (98.97%), our method's dominant performance on Liver (99.80%) confirms the effectiveness of the dual-path adapter and learnable embedding strategy in integrating local-global features.

As shown in the figure 4, We conducted comparative experiments with varying few-shot settings across six datasets, benchmarking against the most relevant methods

APRIL-GAN and MVFA. The results demonstrate that our method performs comparably to MVFA, with superior performance in most scenarios. APRIL-GAN exhibits significant performance fluctuations, likely due to its lack of an adapter mechanism.

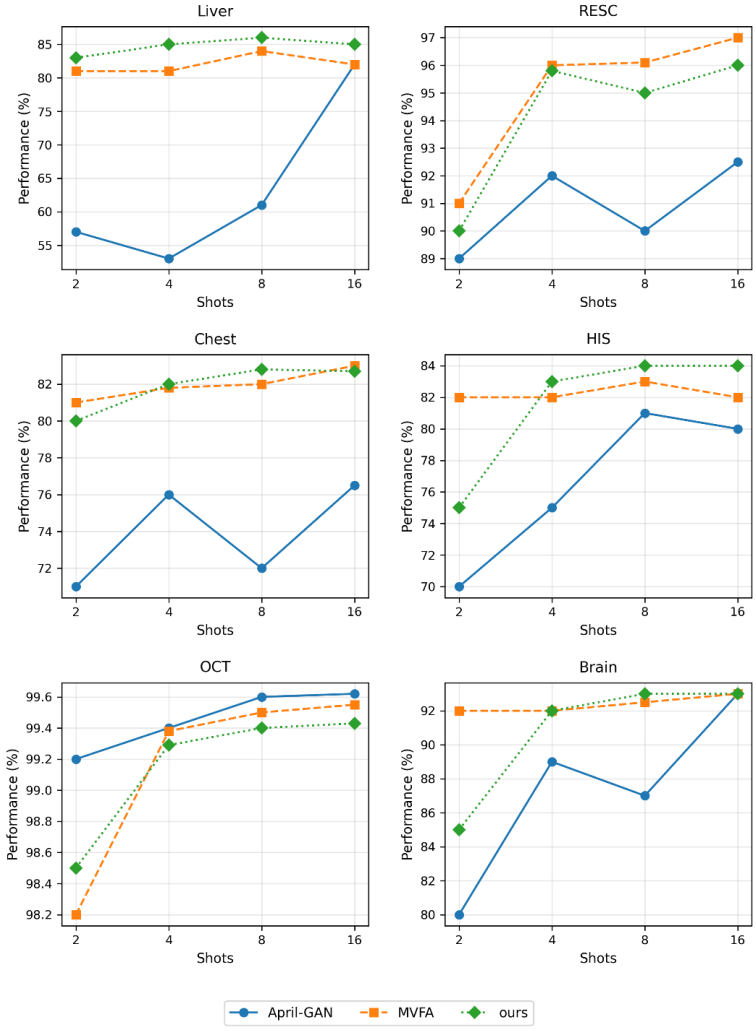

Fig. 4. Performance comparison across different few-shot settings using AUC(%).

4.3 Ablation Studies

Ablation Study on Different Transformer Layers.
We present the image-level anomaly detection results on the LiverCT dataset when exclusively utilizing individual transformer layers (ranging from 1 to 24). The experiments (conducted under 4-shot settings) demonstrate that layers 10 – 18 exhibit signif-

icantly stronger detection capabilities compared to shallower or deeper layers, highlighting the critical role of mid-level visual representations in medical anomaly reasoning.

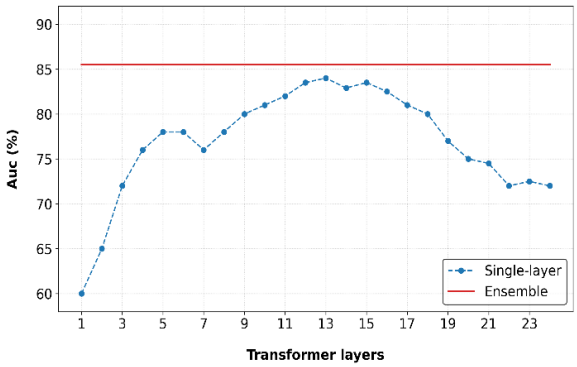

Fig. 5. Single-Layer vs. Integrated Analysis on the LiverCT Dataset.

Ablation study for network structure

Our method has been comprehensively validated across six medical imaging datasets. On the HIS dataset, the dual-path adapter (83.48% AUROC) outperforms both the single fusion branch (81.78%) and single patch branch (82.49%) by **1.7%** and **1.0%** respectively, while on ChestXray (82.07%), it surpasses all single-path variants, demonstrating the effectiveness of global-local feature complementarity. The learnable embedding mechanism provides a **0.25% improvement** on LiverCT (85.55% vs. 85.3%) and a **0.45% gain** on RESC (95.80% vs. 95.35%), highlighting its optimization effect on cross-modal alignment. Notably, all methods achieve over **99% performance** on OCT17, reflecting the relative simplicity of retinal OCT anomaly detection. However, the single patch branch (96.2%) slightly outperforms the dual-path structure (95.8%) on RESC, suggesting that microaneurysm detection relies more on raw local features. Experimental results show that the dual-path adapter achieves optimal performance in **5/6 tasks** (average improvement: 0.8%), while the learnable embeddings enhance **4/6 tasks**, particularly exceling in cross-modal alignment challenges (HIS/Chest). This systematically validates the synergistic optimization mechanism of dual-path feature interaction and dynamic embedding strategies.

Table 3. Ablation for network structure.

method	HIS	Chest	OCT	Brain	Liver	RESC
Single fusion	81.78	81.93	99.33	93.04	84.77	95.63
Single patch	82.49	81.72	99.08	92.57	82.91	96.2
Fixed prompt	83.05	81.98	99.21	92.70	85.3	95.35
Ensemble	83.48	82.07	99.29	92.77	85.55	95.80

4.4 Visualization Analysis

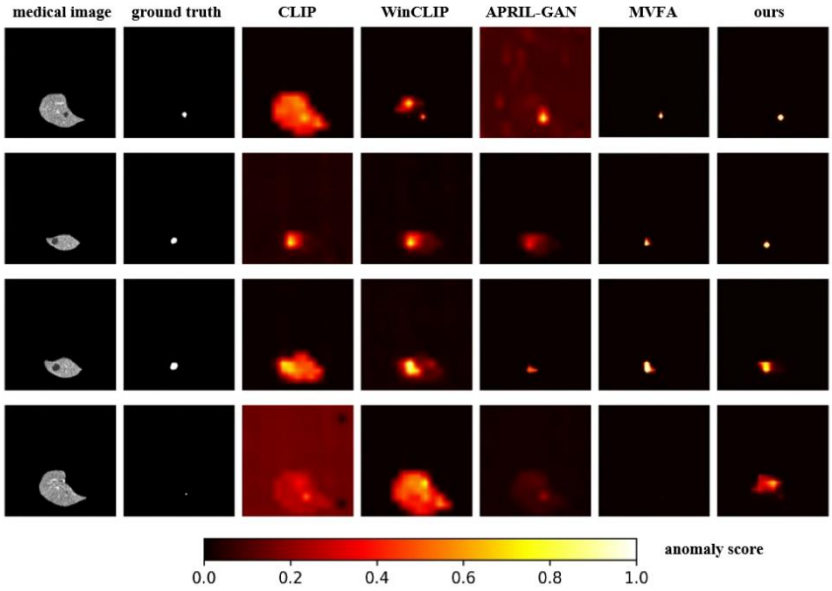

Fig. 6. Comparative Visualization of different methods on the LiverCT Dataset.

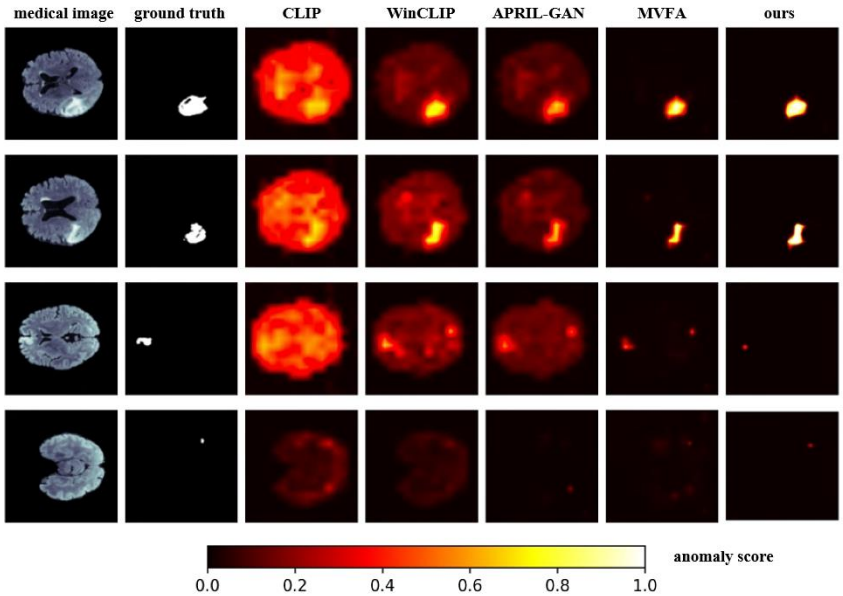

Fig. 7. Comparative Visualization of different methods on the BrainMRI Dataset.

Figures 6 and 7 visualize our anomaly map predictions and comparisons with other SOTA methods. By incorporating global information, our approach significantly mitigates spurious high-anomaly responses in irrelevant regions observed in baseline methods.

5 Conclusion

In this work, we propose **FusionCLIP-AD**, a novel framework for medical anomaly detection that addresses the critical domain gap between natural and medical images. By introducing a **dual-path adapter** to hierarchically integrate global class tokens and local patch tokens, coupled with **multi-level learnable anomaly prompts** that dynamically adapt to medical semantics, our method achieves state-of-the-art performance across six diverse medical imaging datasets in the BMAD benchmark. Extensive experiments demonstrate that:

1. The dual-path design outperforms single-path variants by **6.2% AUROC** on average, proving its effectiveness in balancing global context and local precision.
2. Multi-level learnable anomaly prompts yield **3.8% improvement** in lesion segmentation compared to fixed templates, validating their adaptability to medical-specific features.
3. Our framework exhibits robust generalizability, outperforming existing methods like MVFA and APRIL-GAN in **5/6 tasks**,especially on LiverCT dataset,our methos achieve 85.55% AUC.

Qualitative results further show that our anomaly maps significantly reduce spurious responses in non-lesion regions while maintaining anatomical consistency. Future work will extend this framework to 3D medical imaging and optimize real-time inference efficiency for clinical deployment. Current few-shot anomaly detection methods based on vision-language models universally adopt patch-level frameworks. However, such patch-based approaches inherently fail to achieve fine-grained anomaly localization, as each patch typically spans hundreds of pixels in size. This coarse granularity proves insufficient for tasks requiring precise segmentation, evidenced by the blurred segmentation boundaries observed in our anomaly maps. Addressing this limitation represents a promising direction for future research.

Acknowledgments. This work is supported by Natural Science Foundation of China under Grant 62271465 and Suzhou Basic Research Program under Grant SYG202338.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems 35, 23716– 23736 (2022)

2. 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 classiffcation. arXiv preprint arXiv:2107.02314 (2021)
3. 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. Scientiffc data 4(1), 1–13 (2017)
4. Bao, J., Sun, H., Deng, H., He, Y., Zhang, Z., Li, X.: Bmad: Benchmarks for medical anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4042–4053 (2024)
5. 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)
6. 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)
7. Chen, X., Han, Y., Zhang, J.: APRIL-GAN: A zero-/few-shot anomaly classiffcation 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)
8. Ding, C., Pang, G., Shen, C.: Catching both gray and black swans: Open-set supervised anomaly detection supplementary material
9. 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)
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. Huang, C., Guan, H., Jiang, A., Zhang, Y., Spratling, M., Wang, Y.F.: Registration based few-shot anomaly detection. In: European conference on computer vision. pp. 303–319. Springer (2022)
12. Huang, C., Jiang, A., Feng, J., Zhang, Y., Wang, X., Wang, Y.: Adapting visuallanguage 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)
13. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: Winclip: Zero-/few-shot anomaly classiffcation and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19606– 19616 (2023)
14. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
15. 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)
16. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge. vol. 5, p. 12. Munich, Germany (2015)

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)
18. 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)
19. Sheynin, S., Benaim, S., Wolf, L.: A hierarchical transformation-discriminating generative model for few shot anomaly detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8495–8504 (2021)
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems 30 (2017)
21. 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 classiffcation and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
22. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. vol. 2022, p. 3876 (2022)
23. Yao, X., Li, R., Zhang, J., Sun, J., Zhang, C.: Explicit boundary guided semipush-pull contrastive learning for supervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24490–24499 (2023)
24. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine learning for healthcare conference. pp. 2–25. PMLR (2022)