



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Frequency-Aware Prototype Learning in VLM for Intracranial Aneurysm HR-VWI Segmentation

Yixin Wang^{1,2†}, Haipeng Wang^{2†}, Xin Feng³, Xiangyun Liao^{2(✉)}, and Weixin Si^{2(✉)}

¹Southern University of Science and Technology, China

²Shenzhen University of Advanced Technology, China

³Zhujiang Hospital, China

siweixin@suat-sz.edu.cn, xyunliao@gmail.com

Abstract. Intracranial aneurysms are prone to rupture, making accurate delineation crucial for rupture-risk stratification and treatment planning. While CTA is routinely used for lumen geometry, high-resolution vessel wall imaging (HR-VWI) directly visualizes aneurysm wall enhancement (AWE), providing a wall-focused biomarker that supports wall evaluation and rupture-risk assessment. However, aneurysm borders on HR-VWI are often ill-defined and easily confounded by imperfect blood suppression and parent-vessel distraction, thereby obscuring true boundary cues and compromising precise delineation. Vision-language models (VLMs) can introduce semantic priors via text guidance, yet naive dense fusion is frequently dominated by background structures, diluting weak aneurysm cues. To address these challenges, we propose Frequency-Aware Prototype Learning in VLM for intracranial aneurysm HR-VWI segmentation. Our method leverages discrete wavelet transform to decompose multi-scale features and emphasize boundary-related high-frequency cues, coupled with text-conditioned filtering to retain lesion-relevant frequency responses for query updates. Concurrently, we summarize multi-scale representations into compact prototypes, enabling selective interactions between text-guided queries and foreground–context prototypes to suppress vascular distractions and sharpen delineation. Experiments on an in-house HR-VWI aneurysm benchmark demonstrate consistent improvements over strong baselines, achieving 79.18% Dice. Size-stratified evaluation further confirms robust performance across aneurysm scales, with particularly clear gains for small lesions, supporting more reliable aneurysm quantification for clinical decision-making. Code and test set will be released at <https://github.com/wwyx2722-boop/FAPL-VLM>.

Keywords: HR-VWI aneurysm segmentation · Vision–language model · Frequency-aware enhancement · Prototype learning

[†] Equal contribution.

(✉) Corresponding authors.

1 Introduction

Intracranial aneurysm (IA) rupture can lead to catastrophic subarachnoid hemorrhage without clear prodromal signs, representing a major cause of hemorrhagic stroke, so early detection and longitudinal follow-up in high-risk populations are clinically critical [7]. Digital subtraction angiography (DSA) is the reference standard for IA assessment and has enabled deep-learning diagnosis on 3D DSA [13], yet its invasiveness and cost limit large-scale screening, motivating non-invasive modalities such as CTA and MRA. For reliable diagnosis, risk stratification, and treatment planning, accurate aneurysm segmentation is crucial, as it directly affects the precision of size and morphology and the repeatability of progression assessment across follow-ups.

Beyond lumen-centric angiography, high-resolution vessel wall imaging (HR-VWI) directly depicts aneurysm wall enhancement (AWE), a clinical biomarker of inflammation-mediated wall degeneration linked to aneurysm instability and rupture risk. However, IAs are small targets in HR-VWI, where weak boundaries and slow-flow confounders obscure wall signals, attenuating frequency cues and aggravating class imbalance. Stable and precise aneurysm (and wall) segmentation on HR-VWI would enable enhancement grading and signal quantification, allowing joint analysis with morphologic and hemodynamic indicators to clarify whether rupture risk is driven more by morphology, flow impingement, or inflammation-mediated wall degeneration, thereby supporting interpretable, individualized intervention decisions.

Despite progress in IA segmentation (e.g., semi-supervised CTA [10]), automatic IA segmentation remains a formidable challenge. As shown in Fig. 1, CTA offers clear lumen geometry but limited wall characterization, while HR-VWI captures wall signal for rupture-risk stratification yet presents weak, artifact-prone boundaries, where subtle discontinuities are critical [20]. Existing methods face coupled issues: CNNs may blur fine boundaries under imbalance, and dense global attention are often dominated by background/parent vessels, diluting lesion cues [12]. Meanwhile, frequency-domain boundary cues are often underutilized and can be attenuated by downsampling and aggregation [5].

Motivated by these insights, we propose a language-guided IA segmentation framework that integrates frequency-aware feature refinement with prototype-guided dynamic attention to address low-contrast boundaries and severe class imbalance, which enhances boundary delineation while suppressing background and parent vessel interference for robust segmentation. Our main contributions are summarized as follows:

1. We introduce a frequency-aware strategy to better capture boundary cues under low contrast, reducing boundary leakage.
2. We propose prototype-guided dynamic attention to improve lesion focusing and robustness under extreme foreground sparsity.
3. We present, to our knowledge, the first HR-VWI intracranial aneurysm dataset for segmentation and validate the proposed method across multiple datasets, laying the foundation for HR-VWI-based enhancement grading.

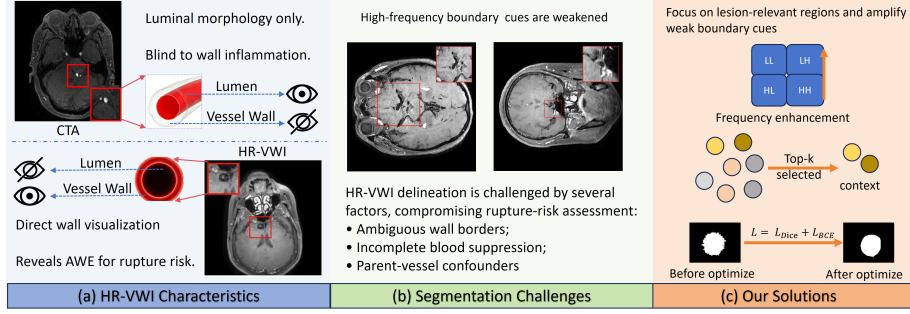


Fig. 1. Motivation and overview. (a) CTA depicts lumen morphology only, whereas HR-VWI visualizes the vessel wall and aneurysm wall enhancement (AWE) for rupture-risk assessment. (b) ambiguous borders, incomplete suppression, and confounders weaken high-frequency cues and hinder delineation. (c) We enhance frequency features and use prototype-guided focused interactions to suppress distractions for sharper segmentation.

2 Method

Figure 2 illustrates the proposed framework, which uses a SEEM[22] visual encoder and a language encoder to initialize learnable segmentation queries. Text-guided Frequency Enhancement Modul(TFEM) injects language-consistent high-frequency cues to enhance weak boundaries, while Dual Prototype Bank and Contrastive Gating Mechanism(DPBM) leverages foreground and context prototypes for gated, lesion-focused query refinement. The refined queries are decoded by the mask projection head to produce segmentation masks.

2.1 Dual Prototype Bank and Contrastive Gating Mechanism (DPBM)

Complex anatomical backgrounds (e.g., vessel textures and organ structures) may induce semantically irrelevant activations and distract segmentation queries from lesion regions. To address this issue, we propose the DPBM to distill a compact prototype set from multi-scale visual features, then builds *foreground* and *context* prototype banks according to text relevance [15,16]. The two banks are subsequently used to perform contrastively gated query refinement, encouraging lesion-focused aggregation while suppressing background interference.

Dual prototype bank construction. Given multi-scale features $\{X^I\}_{I=1}^L$, we obtain a candidate prototype set $P_{\text{raw}} = \{p_k\}_{k=1}^{K_{\text{all}}}$ via lightweight multi-scale distillation. With text embeddings T , we pool a global text representation $\bar{t}$ and compute prototype-text relevance $a_k = \cos(p_k, \bar{t})$. Foreground and context banks are formed by selecting the most and least relevant prototypes:

$$P_{\text{fg}} = \text{TopK}(P_{\text{raw}}, a_k; K_{\text{fg}}), \quad P_{\text{ctx}} = \text{BottomK}(P_{\text{raw}}, a_k; K_{\text{ctx}}). \quad (1)$$

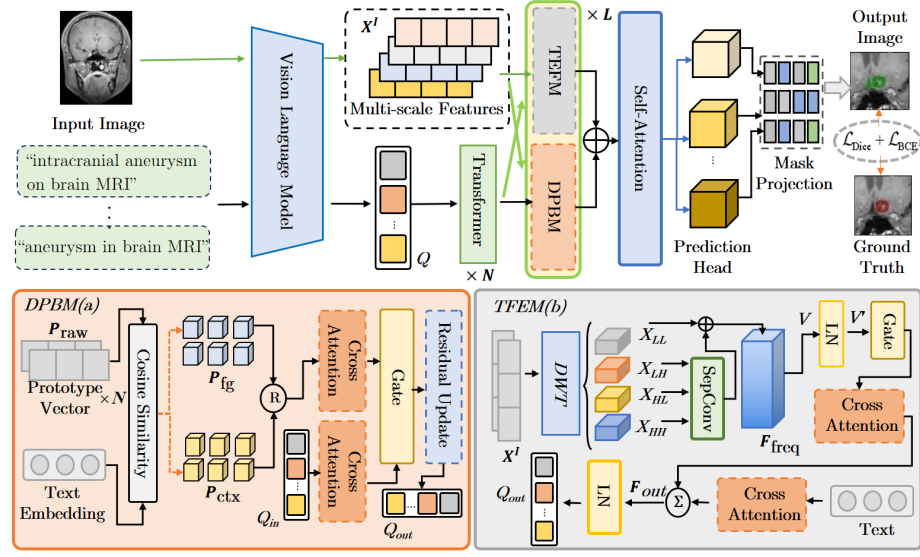


Fig. 2. Overview of the proposed framework for intracranial aneurysm segmentation, which combines multi-scale visual features with text-guided queries for mask prediction. (a) DPBM distills compact aneurysm prototypes and guides dynamic top- k lesion-focused attention, thereby concentrating learning on lesion-relevant regions and reducing interference from background and parent vessels. (b) TFEM performs frequency-aware enhancement to amplify weak boundary cues under low-contrast conditions, improving the sharpness and completeness of the final segmentation. Q is the query; K/V are P_{fg}/P_{ctx} (DPBM) or F_{out} (TFEM). MHCA: multi-head cross-attention; “cos”: cosine alignment.

Contrastive gated query refinement. Given segmentation queries Q , we attend to P_{fg} and P_{ctx} separately and combine the two updates with a contrastive gate g , which is predicted from pooled bank descriptors. The final update is:

$$Q_{out} = Q + g \odot \text{MHCA}(Q, P_{fg}, P_{fg}) - (1 - g) \odot \text{MHCA}(Q, P_{ctx}, P_{ctx}). \quad (2)$$

This gated contrast encourages injection of lesion-relevant prototypes and suppresses context leakage, improving query localization under complex anatomical backgrounds.

2.2 Text-guided Frequency Enhancement Module (TFEM)

High-frequency responses preserve fine-grained textures and boundaries but are more sensitive to artifacts and perturbations [18,6]. To more effectively leverage boundary cues in HR-VWI, we propose TFEM to enhance frequency details and suppresses semantically irrelevant activations under language guidance, and then injects the filtered cues into segmentation decoding.

Frequency enhancement. Given a feature map X , TFEM performs a Haar-based DWT to separate low- and high-frequency components. High-frequency bands are enhanced via a lightweight depthwise separable convolution and channel-wise gate to amplify boundaries while suppressing noise. Then the low- and high-frequency features are fused into a frequency-enhanced representation F_{freq} .

Language-guided filtering and query injection. We obtain an updated token sequence via cross-modal interaction[17] and compute token-wise visual-text relevance using the maximum cosine similarity over text tokens [4]. A soft gate then modulates a residual update to filter F_{freq} :

$$F_{\text{out}} = V + g \odot (V' - V). \quad (3)$$

Finally, the filtered frequency representation is injected into segmentation queries via cross-attention to suppress noisy high-frequency responses:

$$Q_{\text{out}} = \text{LN}(Q + \text{MHCA}(Q, F_{\text{out}}, F_{\text{out}})). \quad (4)$$

3 Experiment

3.1 Dataset

We introduce HiRes-Aneurysm (HiRA), the first HR-VWI intracranial aneurysm dataset, comprising contrast-enhanced HR-VWI scans from 83 patients (2,182 images in total). HiRA was approved by the local IRB/Ethics Committee; an expanded dataset is planned for future work. In contrast to lumen-focused angiographic imaging, HR-VWI provides vessel-wall-oriented visualization that specifically captures AWE, offering improved wall-background contrast and facilitating more complete boundary delineation for clinical evaluation. For voxel-level ground truth, aneurysm masks were independently annotated by two physicians, with any discrepancies resolved by a third senior physician. To evaluate performance across lesion scales, we further stratified HiRA into three size-specific subsets: HiRes-Aneurysm-S (< 7 mm), HiRes-Aneurysm-M (7–10 mm), and HiRes-Aneurysm-L (> 10 mm), comprising 266, 1,109, and 807 images, respectively. In addition, each image is paired with a short radiology-style prompt (e.g., “intracranial aneurysm on brain MRI”) for alignment and classification.

To assess generalizability beyond aneurysm segmentation, we additionally evaluate on two Medical Segmentation Decathlon (MSD) datasets, MSD Lung and MSD Pancreas [14], whose ambiguous boundaries and complex surrounding context are structurally analogous to HR-VWI aneurysm delineation.

3.2 Implementation Details

We implemented our method in PyTorch, training on two NVIDIA RTX 4090 GPUs with a FocalNet-Large visual backbone. The model was optimized using AdamW, with the learning rate decaying from 1×10^{-4} to 1×10^{-6} via a cosine

annealing schedule. To ensure fair comparison, images were resized to a uniform resolution of 1024×1024 with a batch size of 4, and the Interaction Module’s hidden dimension was set to 256. In terms of efficiency, our model has 215M parameters and 1.236 TFLOPs, compared with 182.16M parameters and 3.29 TFLOPs for CausalCLIPSeg, making it feasible for practical HR-VWI aneurysm analysis.

3.3 Evaluation and Interpretability Assessment

We compared our method against uni-modal baselines without text guidance (U-Net++[21], nnU-Net[9], TransUNet[2], Swin-UNet[1]) and recent multi-modal text-guided approaches (LGA[8], CausalCLIPSeg [3], LViT [11], FMISeg [19]). Across all benchmarks, text-guided models consistently outperform those without, indicating that textual priors provide complementary semantic cues.

Table 1. Segmentation performance (%) on HiRA stratified by aneurysm size.

Methods	HiRA Dataset					
	Small (< 7 mm)		Medium (7-10 mm)		Large (> 10 mm)	
	Dice	IoU	Dice	IoU	Dice	IoU
nnU-Net[9]	62.78	50.53	73.17	61.91	76.31	66.87
U-Net++[21]	60.77	49.59	70.62	61.33	75.60	66.19
TransUNet[2]	47.54	36.06	63.57	49.88	75.21	62.89
Swin-UNet[1]	59.63	47.64	69.99	59.86	75.96	63.35
LGA[8]	61.13	48.74	67.49	58.58	75.03	63.02
CausalCLIPSeg[3]	56.67	43.72	58.02	43.97	74.02	62.52
LViT[11]	66.75	55.88	74.59	65.92	76.17	67.05
FMISeg[19]	67.59	55.69	74.17	62.27	77.86	65.74
Ours	68.16	56.26	75.07	63.12	79.18	67.45

As shown in Table 1, our method achieves the best Dice across all three size subsets, obtaining 68.16, 75.07, and 79.18 Dice on small (< 7 mm), medium (7–10 mm), and large (> 10 mm) aneurysms, respectively. The gains are most pronounced on small aneurysms, the most challenging regime due to weak boundaries and extreme foreground sparsity, where our method surpasses the strongest text-guided baseline while remaining competitive in IoU. This indicates that the frequency-aware enhancement and prototype-guided attention jointly improve robustness across lesion scales rather than only on easy, large targets. The qualitative comparisons in Fig. 3 further show that our predictions produce sharper, more complete boundaries with fewer false positives on background and parent vessels than competing methods.

On the two MSD datasets (Table 2), our method also achieves the best performance, reaching 71.66/60.84 (Dice/IoU) on Task06 Lung and 63.12/51.84 on Task07 Pancreas.

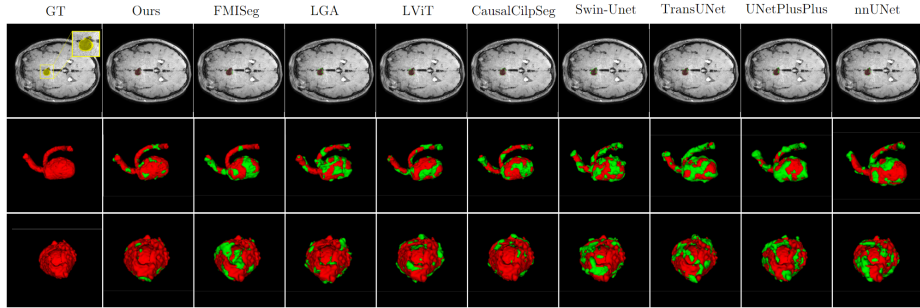


Fig. 3. Representative qualitative results on HiRA. The three rows show slice, aneurysm, and wall segmentation. The first column shows the ground truth (GT), and the remaining columns show predictions. Red indicates predicted masks, and green highlights segmentation errors, including false positives and false negatives. As HiRA is designed for aneurysm analysis, we visualize prediction errors on the aneurysm and vessel-wall surfaces to reveal the spatial relationship between aneurysm segmentation and wall assessment, showing that our method yields more reliable wall and aneurysm delineation for practical HR-VWI analysis.

Table 2. Performance comparison with existing medical image segmentation methods on the Task06 Lung and Task07 Pancreas datasets.

Methods	Backbone	Text	Task06 Lung		Task07 Pancreas	
			Dice	IoU	Dice	IoU
nnU-Net[9]	CNN	×	60.20	49.56	56.40	44.32
U-Net++[21]	CNN	×	57.88	46.19	56.13	44.46
TransUNet[2]	Hybrid	×	60.96	50.35	55.86	43.65
Swin-UNet[1]	Hybrid	×	60.13	49.88	56.58	44.61
LGA[8]	Transformer	✓	68.52	57.54	60.91	48.74
CausalCLIPSeg[3]	Hybrid	✓	68.38	57.78	60.25	48.67
LViT[11]	Hybrid	✓	69.13	59.25	61.70	49.82
FMISeg[19]	CNN	✓	70.41	60.53	62.28	50.89
Ours	Hybrid	✓	71.66	60.84	63.12	51.84

3.4 Ablation Study

We conduct ablation studies and report the results in Table 3. The Baseline performs the worst when both modules are absent. Enabling either TFEM or DPBM consistently improves performance across all datasets, and using both achieves the best overall results, indicating complementary benefits.

Module contribution. Compared with the Baseline, enabling TFEM or DPBM yields consistent gains (about +0.5% ~ +1.3% Dice, with similar IoU improvements), and using both achieves the best performance with Dice increases of

Table 3. Ablation study on HiRA, MSD Task06 Lung, and MSD Task07 Pancreas.

Method	HiRA		Task06 Lung		Task07 Pancreas	
	Dice	IoU	Dice	IoU	Dice	IoU
A. Baseline (without any module)						
Baseline	70.52	58.38	70.26	58.24	62.07	50.44
B. Text-guidance ablation (only one module)						
TFEM <i>w/o</i> Text	70.15	57.86	69.87	57.82	61.75	50.12
TFEM <i>w/</i> Text	71.64	59.60	70.78	58.53	62.88	51.23
DPBM <i>w/o</i> Text	71.24	59.13	70.42	58.25	62.17	50.36
DPBM <i>w/</i> Text	71.77	59.77	70.89	58.67	62.65	50.86
C. Full model (TFEM + DPBM)						
Ours (Both)	72.22	61.37	71.66	60.84	63.12	51.84

+1.70%, +1.40%, and +1.05% on HiRA, Task06 Lung, and Task07 Pancreas, confirming complementary effects. **Text guidance.** Enabling text guidance in only one module at a time consistently boosts TFEM ($\sim +0.9\%$ to $+1.5\%$ Dice) and DPBM ($\sim +0.5\%$ Dice) across datasets, validating the effectiveness of semantic priors. Unlike prior text-guided methods that mainly use text for general conditioning, our aneurysm-specific HR-VWI prompts act as a targeted semantic prior for feature selection, improving Dice by +1.3, +1.1, and +1.5 points for TFEM, DPBM, and the full model, respectively (corrected $p < 0.05$). **Statistical analysis.** Paired Wilcoxon signed-rank tests with Holm–Bonferroni correction confirm the gains are not random: our full model significantly surpasses the Baseline and the best single-module variant on all datasets and metrics, and TFEM *w/* Text surpasses TFEM *w/o* Text on most metrics (all corrected $p < 0.05$).

4 Conclusion

In this work, we introduce the first HR-VWI intracranial aneurysm dataset and propose a language-guided segmentation method addressing two key HR-VWI challenges: extreme foreground imbalance and low-contrast boundaries. The method extracts frequency-domain boundary cues via discrete wavelet transform and uses a prototype-guided, lesion-focused attention mechanism to emphasize lesion regions while suppressing background and vessel-trunk interference. Experiments on multiple datasets show consistent improvements over strong baselines, with enhanced robustness in highly imbalanced, low-contrast settings, particularly for small-target HR-VWI aneurysms.

Limitations. TFEM relies on text guidance to constrain frequency noise: without it (TFEM *w/o* Text), unconstrained high-frequency cues admit artifacts and vessel edges, dropping below the Baseline. This text dependence and the limited

size of HiRA remain to be addressed; our contribution is an effective integration of existing cues rather than a new mechanism.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (No. 62372441), the Guangdong Basic and Applied Basic Research Foundation (No. 2026B1515020093), the Shenzhen Science and Technology Program (Grant No. RCYX20231211090127030), and the Research Program of the Department of Education of Guangdong Province (Grant No. 2025ZDZX3027).

Disclosure of Interests. The authors declare no competing interests.

References

1. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
2. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis* **97**, 103280 (2024)
3. Chen, Y., Wei, M., Zheng, Z., Hu, J., Shi, Y., Xiong, S., Zhu, X.X., Mou, L.: Causalclipseg: Unlocking clip’s potential in referring medical image segmentation with causal intervention. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 77–87. Springer (2024)
4. Cheng, P., Lin, L., Lyu, J., Huang, Y., Luo, W., Tang, X.: Prior: Prototype representation joint learning from medical images and reports. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21361–21371 (2023)
5. Finder, S.E., Amoyal, R., Treister, E., Freifeld, O.: Wavelet convolutions for large receptive fields. In: European conference on computer vision. pp. 363–380. Springer (2024)
6. Guo, Y., Cai, Q.: Bgdiffseg: a fast diffusion model for skin lesion segmentation via boundary enhancement and global recognition guidance. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 150–159. Springer (2024)
7. Hsu, W.C., Meuschke, M., Frangi, A.F., Preim, B., Lawonn, K.: A survey of intracranial aneurysm detection and segmentation. *Medical Image Analysis* **101**, 103493 (2025)
8. Hu, J., Li, Y., Sun, H., Song, Y., Zhang, C., Lin, L., Chen, Y.W.: Lga: A language guide adapter for advancing the sam model’s capabilities in medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 610–620. Springer (2024)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* **18**(2), 203–211 (2021)
10. Li, C.Z., Liu, R.Q., Zhong, H.X., Fan, J.M., Si, W.X., Zhang, M., Heng, P.A.: Semi-supervised intracranial aneurysm segmentation from cta images via weight-perceptual self-ensembling model. *Journal of Computer Science and Technology* **38**(3), 674–685 (2023)

11. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: Lvit: language meets vision transformer in medical image segmentation. *IEEE transactions on medical imaging* **43**(1), 96–107 (2023)
12. Nam, J.H., Syazwany, N.S., Kim, S.J., Lee, S.C.: Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 11480–11491 (2024)
13. Ou, C., Qian, Y., Chong, W., Hou, X., Zhang, M., Zhang, X., Si, W., Duan, C.Z.: A deep learning-based automatic system for intracranial aneurysms diagnosis on three-dimensional digital subtraction angiographic images. *Medical physics* **49**(11), 7038–7053 (2022)
14. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063* (2019)
15. Song, Y., Du, X., Zhang, Y., Xu, C.: Multi-shot prototype contrastive learning and semantic reasoning for medical image segmentation. In: *International conference on medical image computing and computer-assisted intervention*. pp. 578–588. Springer (2023)
16. Tang, S., Yan, S., Qi, X., Gao, J., Ye, M., Zhang, J., Zhu, X.: Few-shot medical image segmentation with high-fidelity prototypes. *Medical Image Analysis* **100**, 103412 (2025)
17. Tian, C., Hu, Y., Zhang, M., Liao, X., Lv, J., Si, W.: Self-prompt contextual learning with axialmamba for multi-label segmentation in carotid ultrasound. *Expert Systems with Applications* **274**, 126749 (2025)
18. Yang, J., Shen, S., Fan, X., Chen, N., Gao, Z., Huang, L., Zhan, Y.: Efms-net: Efficient frequency-enhanced multi-scale network for ischemic stroke segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 161–171. Springer (2025)
19. Yu, B., Yang, J., Du, Z., Huang, Y., Li, C., Wang, L.: Frequency-domain multi-modal fusion for language-guided medical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 278–288. Springer (2025)
20. Zhou, Y., Huang, J., Wang, C., Song, L., Yang, G.: Xnet: Wavelet-based low and high frequency fusion networks for fully-and semi-supervised semantic segmentation of biomedical images. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 21085–21096 (2023)
21. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: *International workshop on deep learning in medical image analysis*. pp. 3–11. Springer (2018)
22. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. *Advances in neural information processing systems* **36**, 19769–19782 (2023)