SAM-MSF^2: SAM finetune-based Multi-Scale Feature Fusion for Retinal Vessel Segmentation

***Abstract*—Medical image segmentation is a significant issue in computer vision. Recently, the Segment Anything Model (SAM) has demonstrated powerful segmentation performance in natural images. Although SAM has been extensively evaluated across various domains, its adaptability to retinal vessel segmentation has not yet been explored. To bridge this research gap, this paper proposes a retinal vessel segmentation method based on fine-tuning SAM and multi-scale feature fusion (MSF^2) strategies. This paper begins by fine-tuning the attention matrix in SAM using a low-rank adaptive strategy, allowing the model to learn medical image features. Subsequently, a cross-attention mechanism is employed in the SAM decoder module to fuse shallow and deep features, capturing vascular information at multiple scales. Finally, a post-refinement network is utilized to enhance the segmentation results of the vessels. Extensive experiments on multiple public datasets demonstrate that SAM-MSF^2 achieves superior retinal vessel segmentation performance across different domain datasets. Moreover, our observations indicate that employing few-shot learning for fine-tuning SAM on these datasets is unnecessary. Code is available at https://anonymous.4open.science/r/SAM-MSFF-DF0B/**

***Index Terms*—Segment Anything, Vessel Segmentation, Medical Imaging, Parameter-Efficient Fine-Tuning**

I. Introduction

The Segment Anything Model (SAM) [15], a groundbreaking large model for image segmentation, excels in generating precise target masks either fully automatically or interactively. It has demonstrated outstanding performance in the semantic segmentation of natural images, attracting widespread attention. However, despite its success in natural image scenarios, SAM's performance in medical imaging remains subpar. Thanks to the extensive application of Parameter-Efficient Fine-Tuning (PEFT) methods, works such as SAMMed [2], AutoSAM [11], and Medical SAM Adapter [27] have been proposed, advancing the application of large segmentation models in medical imaging. SAMMed [2] compiled and organized the largest medical image segmentation dataset to date, utilizing the Adapter method [8] for fine-tuning. However, the input size for SAMMed is limited to 256×256 pixels, significantly restricting its segmentation performance on high-resolution images. AutoSAM [11] froze the encoder's weights and replaced the mask decoder with a prediction head that does not require prompts for training and inference. However, AutoSAM does not fine-tune the encoder, raising doubts about whether the features from natural images can achieve optimal results in the medical domain by merely fine-tuning the decoder. The Medical SAM Adapter [27] introduced the Spatial Depth Transposition (SD-Trans) technique for 2D to 3D adaptation and proposed the Hyper-Prompt Adapter (Hyp-Adpt) for prompt condition adaptation. It employed both serial and parallel adapters for fine-tuning. Nevertheless, this work primarily targets organ and tumor segmentation, leaving the potential of SAM for segmenting fundus vessels unexplored.

The following challenges exist for vessel segmentation using SAM fine-tuning methods: 1) Difficulty in interaction. Most prior work (such as SAM [15], and SAMMed [2]) heavily relies on the quality of prompts. The complex structure of blood vessels is not suitable for this interactive segmentation method, which hinders the application of SAM in vessel segmentation. 2) Poor boundary segmentation. SAM itself handles mask boundary details rather coarsely, whereas vessel segmentation critically depends on accurately delineating boundary details. 3) Input size. The transformer architecture requires a fixed input size. Previous approaches often resize the image directly, resulting in the loss of detail-rich information crucial for vessel segmentation. 4) Mulit-scale nature of vessels. The previous work has not optimized for multi-scale features. Generally, the first-layer features of the image encoder capture more general image boundary details, while the final-layer features contain more global image context information. For thin vessel segmentation, boundary details captured by the first layer are essential; for thick vessel segmentation, global information from the final layer is necessary. SAM-HQ [14] proposes a multi-scale-based post-refinement scheme suitable for fine-tuning models. This method directly fuses the first-layer and final-layer transformer features by addition. However, the importance of these features should vary in different regions, i.e., in areas with thinner vessels, the first-layer features should be emphasized, and vice versa for thicker vessels.

To address the above challenges, this paper proposes a new SAM finetune-based multi-scale feature fusion (SAM-MSF^2) method and applies it to retinal vessel segmentation tasks without prompt setting. Specifically, this paper first fine-tunes the attention matrices in the SAM encoder using a low-rank adaptive (LoRA) strategy [9], enabling the model to learn features from medical images. Subsequently, a cross-attention mechanism is employed to fuse shallow and deep features, capturing vessel information across multiple scales, which are then fed into the SAM decoder module. Finally, a post-refinement network enhances the segmentation of vessels. The main contributions are as follows:

- SAM-MSF^2 represents the first attempt, to our knowledge, to apply SAM to vessel segmentation tasks. We also evaluate the performance of various fine-tuning methods for vessel segmentation.

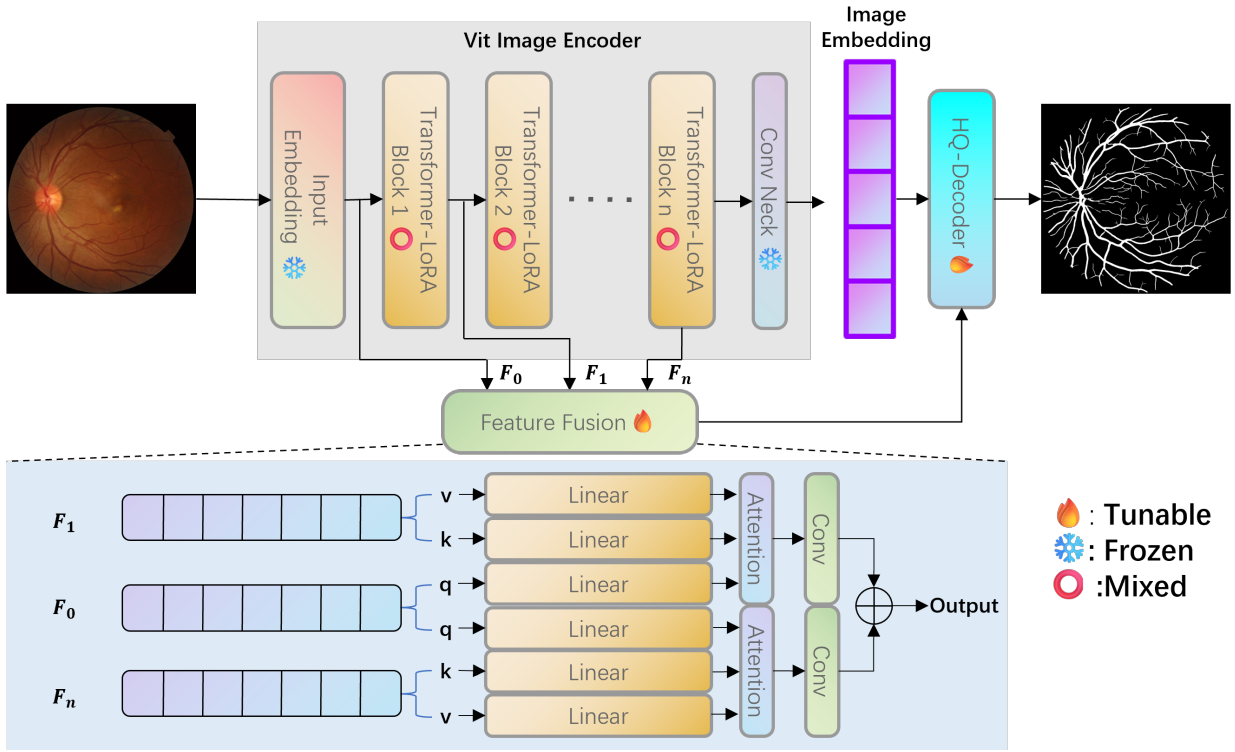

Fig. 1. The architecture of the proposed SAM-MSF^2. It consists of an image encoder with LoRA and a mask decoder with a feature fusion module.

- We introduce a cross-attention method to guide the fusion of feature layers within the image itself. By establishing attention connections between these layers, we enhance the utilization of both local and global information in the image, thereby improving the accuracy and robustness of vessel segmentation.
- Extensive experiments demonstrate that our approach achieves state-of-the-art segmentation results on various publicly available retinal vessel segmentation datasets. Additionally, we observe that fine-tuning SAM is unnecessary for few-shot learning across datasets from different domains.

II. Methods

The framework of our model (SAM-MSF^2) is shown in Fig. 1, consisting of an image encoder with LoRA and a mask decoder with a feature fusion module. To avoid the impact of SAM's original low-resolution input on our task, we divided the image into several patches for segmentation inference. Finally, we stitched the image patches together into a complete image through dilation prediction. In the following sections, we will provide a detailed explanation of how our method fine-tunes the SAM image encoder and integrates multi-scale features for mask decoding.

A. Image Encoder with LoRA

The Image Encoder is the most parameter-intensive component of SAM, and fully tuning this module would incur significant computational costs. To maximize the use of the medical domain knowledge integrated into SAM, we introduced LoRA fine-tuning techniques [9], a classic parameter-efficient fine-tuning method. For a pre-trained weight matrix $W_0 \in R^{d\times k}$. We keep the original weight matrix parameters W_0 unchanged and only fine-tune and update the weight matrix $\triangle W$. The updated weight matrix is decomposed into the product of two low-rank matrices A and B, where $A \in R^{r\times k}, B \in R^{d\times r}$. Its input x will be fed into the original weight matrix W_0 and the updated weight matrix $\triangle W$:

$$Wx = W_0x + \triangle Wx = W_0x + BAx \tag{1}$$

Specifically, we freeze all the parameters in the original image encoder, including the Adapter introduced in SAMMed2d [2], and deploy LoRA matrices for the attention matrices of n Transformer blocks. To balance performance and parameter count, we empirically use rank-16 LoRA matrices to the attention matrices (Q, K, V, and O) rather than add LoRA matrices only to the Q and V matrices [5]. As illustrated in Fig. 2, in the Transformer Block, x is input into the fine-tuned attention module, resulting in the attention output A_i:

$$A_i(x) = W_ix + B_iA_ix, \quad i = q, k, v, o \tag{2}$$

B. Mask Decoder with feature fusion module

Based on the original Mask Decoder, we modify the decoder structure with inspiration from SAM-HQ [14], and introduce cross-attention fusion to enhance multi-scale image perception. Specifically, we extract features from the first and last layers of Transformer-LoRA blocks and compute cross-attention using

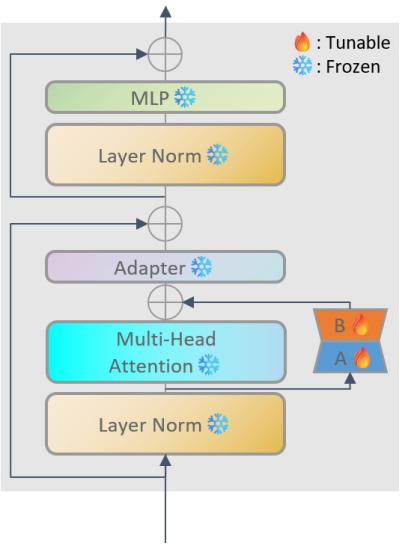

Fig. 2. Transformer-LoRA Block. During fine-tuning, only low-rank matrices A and B are updated.

image embeddings. The outputs after feature fusion F_{fuse} are subsequently input into the SAM-HQ decoder, illustrated in Fig. 1. Specifically, the output F_i of the i-th Transformer-LoRA Block serves as the keys (k) and values (v) in the cross-attention computation, while the image embedding F_0 serves as the queries (q). Then, the shallow features and deep features after attention-weighted calculation are added pixel-wise. This paper takes the first-layer features as shallow features and the last-layer features as deep features.

$$F_{fuse} = Conv(Attention(Q, K_1, V_1)) \oplus \tag{3}$$
$$Conv(Attention(Q, K_n, V_n)) \tag{4}$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

$$Q = W_q F_0, \tag{6}$$
$$K_i = W_k F_i, \quad i = 1, n \tag{7}$$
$$V_i = W_v F_i, \quad i = 1, n \tag{8}$$

Ultimately, SAM-HQ's decoder HQ decodes the segmentation mask using the output F_{img} of the Image Embedding and the output of the feature fusion module F_{fuse}:

$$Mask = HQ(F_{fuse}, F_{img}) \tag{9}$$

C. Training and Inference Strategies

Before training, we crop the images into 256×256 patches to fit the image input size of SAM and remove images where the vessel region in the mask is less than 1%. During the inference phase, we first pad the image dimensions with zero pixels to the nearest multiple of 256. We then perform sliding window prediction using a window size of 256×256 and a stride of 128. For each prediction, only the central 256×256 region is retained for stitching, resulting in the final predicted mask. The loss function supervising the mask predictions is a linear combination of cldice loss [23] and dice loss [17], with a ratio of 1:1. The loss function can be represented as:

$$\mathcal{L} = 0.5\mathcal{L}_{dice} + 0.5\mathcal{L}_{cldice} \tag{10}$$

$$\mathcal{L}_{dice} = 1 - \frac{2 \times |\widehat{Y} \cap Y|}{|\widehat{Y}| + |Y|} \tag{11}$$

$$\mathcal{L}_{cldice} = 1 - 2 \times \frac{T_{prec}(\widehat{Y}_S, Y) \times T_{sens}(Y_S, \widehat{Y})}{T_{prec}(\widehat{Y}_S, Y) + T_{sens}(Y_S, \widehat{Y})} \tag{12}$$

where Y represents the ground truth, $\widehat{Y}$ denotes the predicted value. Y_S and $\widehat{Y}_S$ are the soft skeleton [23] of Y and $\widehat{Y}$, respectively. T_{prec} and T_{sens} refer to precision and sensitivity.

III. Experiments and Results

A. Dataset

We utilize six publicly available datasets for retinal blood vessel segmentation to evaluate the performance of our SAM-MSF^2 model: STARE [7], CHASE_DB1 [6], DRIVE [24], DRIVE2, HRF [18], and FIVES [13]. Here are brief descriptions of each dataset:

- STARE [7]: Contains 20 fundus images, including 10 with lesions and 10 without, with an image resolution of 605 × 700 pixels.
- CHASE_DB1 [6]: Consists of 28 color retina images (999 × 960 pixels each) collected from both eyes of 14 school children.
- DRIVE [24]: Includes 40 color fundus images, with 7 cases featuring abnormal pathology. Each image has a resolution of 584 × 565 pixels.
- DRIVE2: An augmented dataset based on the DRIVE dataset, hosted on Kaggle.
- HRF [18]: Comprises 45 images for retinal vessel segmentation, with image sizes of 3304 × 2336 pixels.
- FIVES [13]: Contains 800 high-resolution multi-disease color fundus photographs, each sized 2048 × 2048 pixels.

Our SAM-MSF^2 uses the training and validation sets from the FIVES dataset for model training.

B. Implementation Details

Our method is implemented in PyTorch and trained on 4 NVIDIA 3090 GPUs, each with 24GB memory. We use the AdamW optimizer with an initial learning rate of 1e-4, with the learning rate divided by 2 at the 5th and 10th epochs. During training, the batch size is 32.

TABLE I
QUANTITATIVE COMPARISON OF THE SEGMENTATION EFFECT OF DIFFERENT METHODS ON SIX DATASETS

Dataset	*Method*	*IoU*↑	*Dice*↑	*clDice*↑	*AUC*↑	*HD95*↓
STARE	Ours	**45.68**	**62.51**	**55.35**	**78.16**	**87.70**
	LoRA+HQ	44.93	61.78	54.56	77.79	91.28
	LoRA	42.40	59.27	51.95	76.50	96.78
	SAMMed2d(Fine-tuned)	44.63	61.67	55.06	78.00	89.14
CHASE_DB1	Ours	**58.02**	**73.38**	**81.33**	**89.69**	**46.04**
	LoRA+HQ	57.58	73.18	81.28	89.58	46.78
	LoRA	56.84	72.41	79.67	88.52	53.34
	SAMMed2d(Fine-tuned)	54.54	70.52	78.86	89.55	54.87
DRIVE	Ours	**50.19**	**66.73**	**60.99**	**80.41**	**46.07**
	LoRA+HQ	50.10	66.64	60.27	80.02	47.41
	LoRA	49.03	65.67	58.04	78.76	51.60
	SAMMed2d(Fine-tuned)	49.22	65.84	58.59	78.65	47.23
DRIVE2	Ours	54.21	70.27	**67.79**	**84.27**	**31.45**
	LoRA+HQ	**54.28**	**70.33**	67.39	84.09	31.97
	LoRA	54.21	70.26	66.13	83.42	34.07
	SAMMed2d(Fine-tuned)	53.86	70.04	66.96	83.18	32.97
HRF	Ours	**64.15**	**78.02**	**78.11**	**89.94**	**74.30**
	LoRA+HQ	63.99	77.90	77.64	89.73	75.32
	LoRA	63.27	77.37	77.27	89.23	81.76
	SAMMed2d(Fine-tuned)	62.9	77.06	77.31	89.15	89.03
FIVES	Ours	**77.33**	**86.28**	**86.65**	**91.20**	**76.79**
	LoRA+HQ	76.63	85.76	85.77	90.91	84.12
	LoRA	75.55	84.97	85.03	90.29	88.33
	SAMMed2d(Fine-tuned)	75.32	84.86	84.97	90.11	91.31

C. Metric

We test the performance of various fine-tuning methods on different datasets. The evaluation indicators selected IoU [29], Dice [17], cldice [23], AUC, [12] and 95% Hausdorff distance(HD95) [22], and their calculations are as follows:

$$IoU = \frac{2 \times |\widehat{Y} \cap Y|}{|\widehat{Y} \cup Y|} \tag{13}$$

$$Dice = \frac{2 \times |\widehat{Y} \cap Y|}{|\widehat{Y}| + |Y|} \tag{14}$$

$$cldice = 2 \times \frac{T_{prec}(\widehat{Y}_S, Y) \times T_{sens}(Y_S, \widehat{Y})}{T_{prec}(\widehat{Y}_S, Y) + T_{sens}(Y_S, \widehat{Y})} \tag{15}$$

$$AUC = \frac{1}{2} \sum_{i=1}^{n-1} (x_{i+1} - x_i) \cdot (y_i + y_{i+1}) \tag{16}$$

where n is the total number of thresholds, x_i represents the false positive rate (FPR) at threshold i, and y_i represents the true positive rate (TPR) at threshold i.

$$HD_{95} = max(h(\widehat{Y}, Y), h(Y, \widehat{Y})) \tag{17}$$

where $h(A, B) = f^{95\%}_{a \in A} \underset{b \in B}{min} ||a - b||$,and $f^{95\%}_{x \in X} g(x)$ denotes the 95% quantile value of $g(x)$ over the set X. [22]

D. Experimental Results

This paper validates the SAM-MSF^2 from two perspectives: evaluating its segmentation performance across different datasets and assessing its few-shot learning capability on various datasets. The models included in the performance comparison are as follows: 1) Fine-tuning the Adapter of SAMMed2d [2] (SAMMed2d (FIne-tuned)); 2) Fine-tuning SAM using LoRA [9] (LoRA); 3) Adding an HQ decoder [14] to model 2) (LoRA+HQ).

We train the models using the training and validation sets from the FIVES dataset and evaluated their segmentation performance and generalization ability across different datasets. Table I presents the segmentation performance of various methods on these datasets. From the table, our method achieves the best or equivalent results across all evaluation metrics compared to existing methods. While it slightly lags behind the LoRA+HQ method on the DRIVE2 dataset, our method demonstrates a significant advantage on all other datasets. This advantage arises from the introduction of cross-attention to process features, allowing for more effective integration of local and global features, resulting in higher quality vascular segmentation masks. Fig. 3 provides a visual comparison of our model's performance with other models on the FIVES test set.

The following experiment involves performing few-shot learning on other datasets using pre-trained weights obtained from training on the FIVES dataset. This experiment aims to

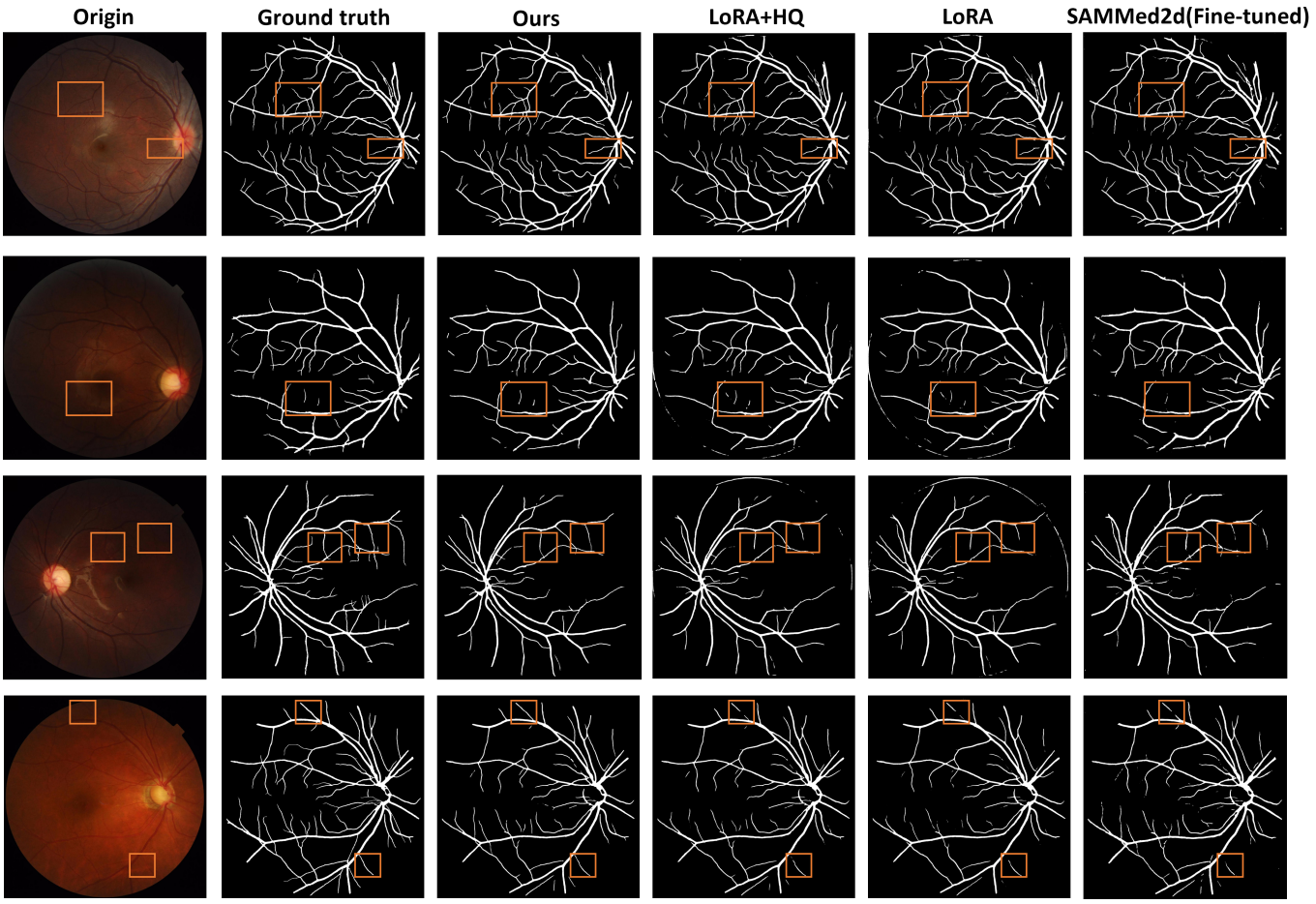

Fig. 3. Comparison of the segmentation results on four samples within FIVES datasets between SAM-MSF[2] and SOTA SAM fine-tuned-based methods.

study the influence of different numbers of training samples on the domain adaptability of fine-tuning SAM. Specifically, we sort the Dice scores of all the images in each dataset in descending order and assign them alternately to the training set and the test set. We then randomly select 20%, 40%, and 100% of the images from the training set for training and evaluate the performance in the test set. To avoid sampling error, the 20% and 40% sample groups are trained five times in different batches and the results are averaged.

Figs. 4 and 5 provide the results of few-shot learning on the CHASE_DB1 and HRF datasets after pre-training various methods on the FIVES dataset. It can be observed that the improvement from few-shot fine-tuning on target datasets is not significant, even with an increased amount of training data. On the CHASE_DB1 dataset, our method achieve a Dice score of 73.32% after training on FIVES. The Dice scores for the 20%, 40%, and 100% groups are 75.27%, 75.56%, and 75.76%, respectively. This means that using just 3 images as the training set improves the score by 1.95%. When the data volume was increased fivefold, the Dice score only improves by an additional 0.49%. On the HRF dataset, our method achieves a Dice score of 77.94% after training on FIVES. The Dice scores for the 20%, 40%, and 100% groups are 78.60%, 78.80%, and 78.98%, respectively. Fine-tuning with 5 images results in a score improvement of only 0.66%, and increasing the data volume fivefold leads to just an additional 0.38% improvement. Therefore, for the vessel segmentation task, the segmentation performance mainly depends on the effectiveness of the pre-trained weights.

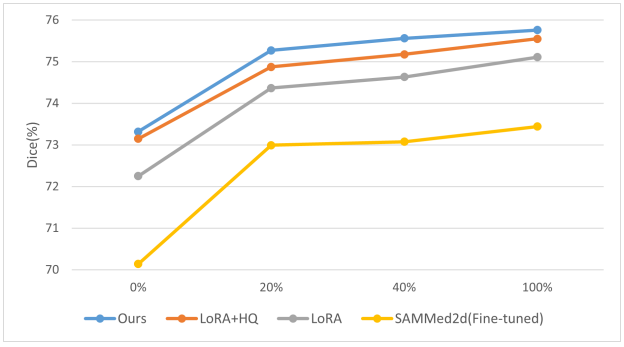

Fig. 4. Improvement of Dice for Different Models under Few-shot Learning in CHASE_DB1 dataset

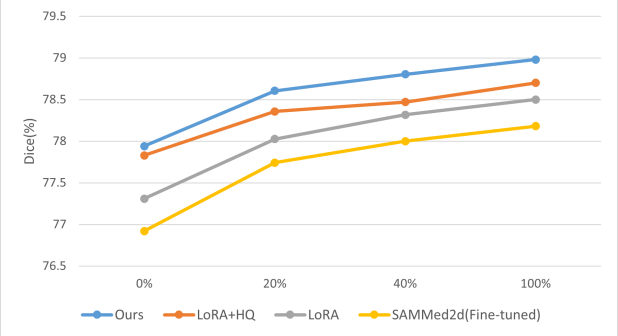

Fig. 5. Improvement of Dice for Different Models under Few-shot Learning in HRF dataset

IV. Discussion

For retinal vessel segmentation tasks, the cost of manual annotation is quite high. Therefore, we aim to focus on the fine-tuning performance of foundational models in the field of retinal segmentation. We have found that the fine-tuned models exhibit better generalization performance on high-resolution datasets compared to other datasets. Retinal vessel segmentation typically requires accurate extraction of fine vessel structures. High-resolution datasets provide richer and more detailed image information, enabling the model to better encode and capture the subtle details of blood vessels. In contrast, low-resolution datasets may face issues of information loss and blurriness, limiting the accuracy of segmentation methods. This also means that large models, with their greater number of parameters, can better learn the general features of blood vessels.

Additionally, we also hope to explore whether large models need extra training to adapt to target domain data. The experimental results did not reflect this necessity. Traditionally, for specific tasks, large models need to be fine-tuned on target domain data to adapt to the unique characteristics of that domain. However, in our experiments, even without additional target domain training, the performance of large models is very close to that after adaptation training. Large models typically possess stronger generalization capabilities, allowing them to learn richer feature representations from large-scale source domain data. This enables large models to better adapt to the diversity and complexity within the target domain.

However, we also need to note that using the patch-based sliding window prediction method, while it can adapt to high-resolution images without loss of accuracy, requires more time during inference. Additionally, high-quality annotated retinal vessel datasets like FIVES are still too scarce to fully demonstrate the advantages of large models. Improving model performance under conditions of data scarcity remains a challenge.

V. Conclusion

We propose a novel SAM-based retinal vessel segmentation method using LoRA fine-tuning and multi-scale feature fusion strategies. This method can optimize the local/global information mismatch that may occur in the feature fusion process of the HQ decoder. We also introduce a cross-attention mechanism that enables the model to decide whether to focus more on local or global information based on image embeddings. We evaluate our approach on six public retinal vascular segmentation datasets, and our approach outperforms other fine-tuning methods on various metrics. Finally, through the exploration of few-shot learning, we demonstrate the potential of the foundational model in the application of retinal vessel segmentation.

References

[1] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic Convolution: Attention over Convolution Kernels," Mar. 2020, arXiv:1912.03458 [cs]. [Online]. Available: http://arxiv.org/abs/1912.03458

[2] J. Cheng, J. Ye, Z. Deng, J. Chen, T. Li, H. Wang, Y. Su, Z. Huang, J. Chen, L. Jiang, H. Sun, J. He, S. Zhang, M. Zhu, and Y. Qiao, "SAM-Med2D," Aug. 2023, arXiv:2308.16184 [cs]. [Online]. Available: http://arxiv.org/abs/2308.16184

[3] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional Feature Fusion," 2021, pp. 3560–3569. [Online]. Available: https://openaccess.thecvf.com/content/WACV2021/html/Dai_Attentional_Feature_Fusion_WACV_2021_paper.html

[4] B. Fazekas, J. Morano, D. Lachinov, G. Aresta, and H. Bogunović, "SAMedOCT: Adapting Segment Anything Model (SAM) for Retinal OCT," Aug. 2023, arXiv:2308.09331 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2308.09331

[5] V. Fomenko, H. Yu, J. Lee, S. Hsieh, and W. Chen, "A note on lora," 2024.

[6] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman, "An ensemble classification-based approach applied to retinal blood vessel segmentation," IEEE Transactions on Biomedical Engineering, vol. 59, no. 9, pp. 2538–2548, 2012.

[7] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," IEEE Transactions on Medical Imaging, vol. 19, no. 3, pp. 203–210, 2000.

[8] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," 2019.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," Oct. 2021, arXiv:2106.09685 [cs]. [Online]. Available: http://arxiv.org/abs/2106.09685

[10] X. Hu, F. Li, D. Samaras, and C. Chen, "Topology-Preserving Deep Image Segmentation," in Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/2d95666e2649fcfc6e3af75e09f5adb9-Abstract.html

[11] X. Hu, X. Xu, and Y. Shi, "How to Efficiently Adapt Large Segmentation Model(SAM) to Medical Images," Jun. 2023, arXiv:2306.13731 [cs]. [Online]. Available: http://arxiv.org/abs/2306.13731

[12] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," IEEE Transactions on knowledge and Data Engineering, vol. 17, no. 3, pp. 299–310, 2005.

[13] K. Jin, X. Huang, J. Zhou, Y. Li, Y. Yan, Y. Sun, Q. Zhang, Y. Wang, and J. Ye, "FIVES: A Fundus Image Dataset for Artificial Intelligence based Vessel Segmentation," Scientific Data, vol. 9, no. 1, p. 475, Aug. 2022, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41597-022-01564-3

[14] L. Ke, M. Ye, M. Danelljan, Y. liu, Y.-W. Tai, C.-K. Tang, and F. Yu, "Segment anything in high quality," in Advances in Neural Information Processing Systems, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 29 914–29 934. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/5f828e38160f31935cfe9f67503ad17c-Paper-Conference.pdf

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment Anything," Apr. 2023, arXiv:2304.02643 [cs]. [Online]. Available: http://arxiv.org/abs/2304.02643

[16] J. Lin, X. Huang, H. Zhou, Y. Wang, and Q. Zhang, "Stimulus-guided adaptive transformer network for retinal blood vessel segmentation in fundus images," Medical Image Analysis, vol. 89, p. 102929, Oct. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523001895

[17] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in 2016 fourth international conference on 3D vision (3DV). Ieee, 2016, pp. 565–571.

[18] J. Odstrcilik, R. Kolar, A. Budai, J. Hornegger, J. Jan, J. Gazarek, T. Kubena, P. Cernosek, O. Svoboda, and E. Angelopoulou, "Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database," IET Image Processing, vol. 7, no. 4, pp. 373–383, Jun. 2013, publisher: John Wiley & Sons, Ltd. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-ipr.2012.0455

[19] S. Pandey, K.-F. Chen, and E. B. Dam, "Comprehensive Multimodal Segmentation in Medical Imaging: Combining YOLOv8 with SAM and HQ-SAM Models," 2023, pp. 2592–2598. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023W/CVAMD/html/Pandey_Comprehensive_Multimodal_Segmentation_in_Medical_Imaging_Combining_YOLOv8_with_SAM_ICCVW_2023_paper.html

[20] J. N. Paranjape, N. G. Nair, S. Sikder, S. S. Vedula, and V. M. Patel, "AdaptiveSAM: Towards Efficient Tuning of SAM for Surgical Scene Segmentation," Aug. 2023, arXiv:2308.03726 [cs]. [Online]. Available: http://arxiv.org/abs/2308.03726

[21] Y. Qi, Y. He, X. Qi, Y. Zhang, and G. Yang, "Dynamic Snake Convolution based on Topological Geometric Constraints for Tubular Structure Segmentation," Aug. 2023, arXiv:2307.08388 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2307.08388

[22] W. Rucklidge, "Locating objects using the hausdorff distance," in Proceedings of IEEE International Conference on Computer Vision, 1995, pp. 457–464.

[23] S. Shit, J. C. Paetzold, A. Sekuboyina, I. Ezhov, A. Unger, A. Zhylka, J. P. W. Pluim, U. Bauer, and B. H. Menze, "clDice – A Novel Topology-Preserving Loss Function for Tubular Structure Segmentation," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2021, pp. 16 555–16 564, arXiv:2003.07311 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2003.07311

[24] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," IEEE Transactions on Medical Imaging, vol. 23, no. 4, pp. 501–509, 2004.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[26] C. Wang, X. Chen, H. Ning, and S. Li, "SAM-OCTA: A Fine-Tuning Strategy for Applying Foundation Model to OCTA Image Segmentation Tasks," Sep. 2023, arXiv:2309.11758 [cs]. [Online]. Available: http://arxiv.org/abs/2309.11758

[27] J. Wu, W. Ji, Y. Liu, H. Fu, M. Xu, Y. Xu, and Y. Jin, "Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation," Dec. 2023, arXiv:2304.12620 [cs]. [Online]. Available: http://arxiv.org/abs/2304.12620

[28] A. Xiao, W. Xuan, H. Qi, Y. Xing, R. Ren, X. Zhang, and S. Lu, "CAT-SAM: Conditional Tuning Network for Few-Shot Adaptation of Segmentation Anything Model," Feb. 2024, arXiv:2402.03631 [cs]. [Online]. Available: http://arxiv.org/abs/2402.03631

[29] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in Proceedings of the 24th ACM international conference on Multimedia, ser. MM '16. ACM, Oct. 2016. [Online]. Available: http://dx.doi.org/10.1145/2964284.2967274

[30] K. Zhang and D. Liu, "Customized Segment Anything Model for Medical Image Segmentation," Oct. 2023, arXiv:2304.13785 [cs]. [Online]. Available: http://arxiv.org/abs/2304.13785