Multimodal Prompt Sequence Learning for Interactive Segmentation of Vascular Structures *

Jongsoo Lim[0009−0005−1659−4231] and Soochahn Lee[0000−0002−2975−2519]

Kookmin University, Seoul, Korea
sclee@kookmin.ac.kr

Abstract. Interactive segmentation tools are necessary to achieve the desired segmentation accuracy for complex target structures, such as vessels in medical images. But existing interactive methods–including those pre-trained on large internet-scale datasets–offer limited mechanisms for users to provide prompts that effectively control segmentation outcomes. In particular, one-at-a-time point or text prompts are often insufficient for correcting errors in vascular segmentation masks. To address these limitations, we propose a novel interactive medical image segmentation method tailored for complex vascular structures. Our approach learns to interpret sequences of multimodal prompts–combining both text and point inputs. By enabling dual mode prompting, the method allows users to add semantic meaning to point-based interactions. Furthermore, by learning from aggregated sequences of prompts, the method captures inter-prompt relationships, enhancing its understanding and response to user input. Quantitative evaluations on six vascular datasets demonstrate that our method outperforms existing approaches. Additionally, it avoids critical failure cases and consistently generates improved segmentation masks across diverse imaging modalities and vascular anatomies.

Keywords: Multimodal Prompt · Sequence Learning · Interactive Segmentation · Vessel Segmentation

1 Introduction

To address inevitable errors in automatic segmentation methods [19, 2, 8, 23], interactive approaches [21, 3, 15] incorporate user input to maintain robust accuracy under diverse conditions. Recently, the Segment Anything Model (SAM) [11] has gained attention due to its strong generalization performance. SAM can be conditioned on various types of prompts, including points and masks, and has been adopted as a foundational model in many downstream methods [9, 16].

However, SAM can be limited in accurately segmenting complex and slender structures, such as vessels in medical images. Due to the intricate and fine-grained nature of vascular structures, traditional point-based prompts may fail

* This work was supported by the National Research Foundation of Korea (NRF) grant (RS-2025-00521972) and Institute of Information & Communications Technology Planning & Evaluation(IITP) grant (No.RS-2025-02219317, AI Star Fellowship (Kookmin University)) funded by the Korea government (MSIT).

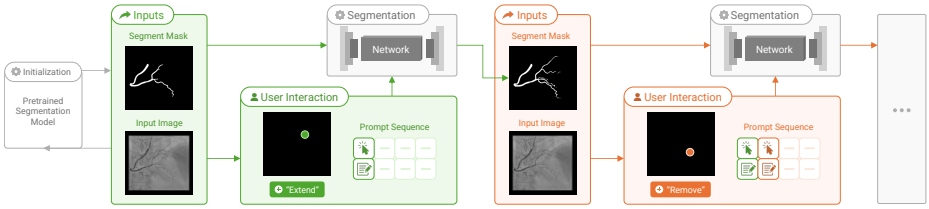

Fig. 1. Overview of the proposed method. The model incorporates sequences of multimodal text and point prompts to enable interactive segmentation of complex vessel structures, providing enhanced user control throughout the process.

to convey sufficient information. As a result, the desired accuracy may not be achieved, even when numerous specific point prompts are provided.

In interactive segmentation, subsequent prompts are often correlated. For instance, users may provide multiple prompts to iteratively refine segmentation in ambiguous or challenging regions. A model that can learn the relationships among sequential prompts is better positioned to capture user intent and enhance segmentation performance.

To this end, we propose an interactive medical image segmentation framework that integrates multimodal prompts with sequential learning. The main contributions are as follows:

1. We propose a novel interactive medical image segmentation method for complex vascular structures that leverages multimodal prompts–specifically, text and point inputs–built on top of the SAM framework. We define a set of task-specific text prompts (e.g., "remove" or "extend") and introduce a process for generating synthetic interactive segmentation sequences for training. While Liu et al. [14] introduced the idea of combining point and text prompts, they did not address learning from sequences of such multimodal inputs.
2. To accommodate the iterative nature of interactive segmentation, we introduce a model architecture and learning strategy that support sequential prompt integration. This enables the model to retain and utilize information from previously provided prompts, allowing accuracy to improve progressively during iterative interactions.
3. We conduct extensive experimental evaluations on multiple vessel segmentation datasets spanning various modalities, including retinal fundus [1, 5, 6], optical coherence tomography angiography (OCTA) [13], and coronary x-ray angiography [17]. Results demonstrate that our method enables more flexible user interaction by leveraging richer information in multimodal prompt sequences, and achieves improved segmentation accuracy of complex vascular structures compared to existing interactive methods [21, 3, 15].

A visual overview of the proposed framework is shown in Fig. 1. We assume the availability of an initial segmentation mask along with the input image, which may be generated by a fully automatic vessel segmentation model. Given this image–mask pair, the user iteratively provides both point and text prompts

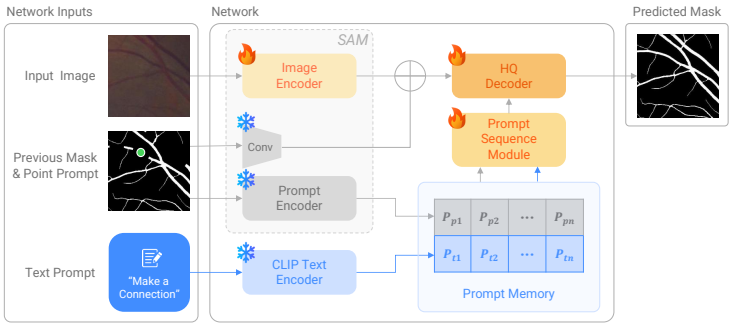

Fig. 2. Model structure of the proposed method. Mask and point prompt encoders from SAM [11], together with the text prompt encoder from CLIP [18] are adopted to enable multimodal prompts. We propose a prompt sequence model (PSM) to combine the sequence of multimodal prompts, which conditions the output segmentation, predicted from the decoder of the SAM-HQ [9]. During training, the encoders are kept frozen to maintain generalizability.

to guide the segmentation model in correcting errors. This interactive process continues until the segmentation mask reaches a satisfactory level of accuracy.

2 Method

2.1 Segmentation Model Architecture

The proposed segmentation network comprises the backbone encoder, text encoder, the prompt sequence module, and the HQ decoder, as illustrated in Fig. 2.

For the backbone encoder, we adopt the components from SAM [11], which include a Vision Transformer (ViT)-based image encoder [4], along with encoder modules to process the point and mask inputs.

For the text prompt encoder, we utilize the CLIP [18] encoder, as a text encoder is not publicly available within the original SAM framework. To ensure compatibility, we append an additional linear layer to the CLIP encoder output to match the dimensionality of the point prompt encoder's latent representation.

The prompt sequence module (PSM) consists of a few simple components for aggregating and encoding prompt sets. It includes either input overlay or feature concatenation mechanisms for prompt aggregation, followed by prompt encoding layers such as LSTM [7] or self-attention [22]. Given the variety of possible configurations for the PSM, we present an ablation study in Sec. 3.4 to analyze their impact.

For the decoder, we adopt the architecture of SAM-HQ [9] to better capture the fine, high-resolution details of vascular structures. The encoded features from the image and the current segmentation mask are combined, and the decoder is conditioned on the encoded multimodal prompts from the PSM to produce the updated mask prediction.

2.2 Training Process

Generating Training Sequence Data We define a training data instance at step k as $d^k = \{I, M^k, p^k, t^k\}$, where I denotes the input image, and M^k, p^k, and t^k represent the segmentation mask, point prompt, and text prompt, respectively. An interactive segmentation sequence is defined as $\mathcal{D} = \{d^0, ..., d^{K-1}\}$, where K is the total number of interaction steps in the sequence. Input images for training can be sourced from various public datasets [1, 5, 13], and an initial segmentation mask M^0 can be generated using a fully supervised model, such as nnUnet [8]. While users can provide t^k and p^k during inference, collecting large numbers of such prompt sequences for training is costly. Although p^k can be generated by identifying erroneous regions in M^k using the GT mask M^{GT}, generating corresponding text prompts t^k is considerably more challenging.

We address this challenge by training a text prompt predictor (TPP) on a sample dataset with predefined text prompt classes. Given that the target regions correspond to vessels, we define a set of five representative text prompt types: $t^{cls} \in$ {"make thinner", "make thicker", "make a connection", "extend", "remove"}. For each class, we follow a class-specific procedure to generate training samples, as detailed below:

- Make thinner (thicker): Skeletonize M^{GT} [12] and estimate the radius via cubic Hermite spline fitting. Then, generate modified masks M^{thin} (M^{thick}) by reconstructing the vessels with reduced (increased) radii.
- Connect: Detect bifurcations on the vessel skeleton using the hit-or-miss transform [20], randomly remove segments near them to create disconnections, and reconstruct the mask M^{conn} by restoring vessel radii.
- Extend: Randomly remove a branch from the vessel skeleton and reconstruct the mask M^{ext} by restoring vessel radii along the modified skeleton.
- Remove: Generate noisy masks M^{rem} by controlling the binarization threshold of M^0 so that the amount of false positive vessel pixels are increased.

Training the Network The TPP takes $\{I, M^{cls}, p^{cls}\}$ as input and predicts the corresponding t^{cls}, where the superscript $cls \in \{thin,\ thick,\ conn,\ ext,\ rem\}$ denotes the prompt classes described above. For the network, we reuse the backbone encoders for image, mask, and point prompts from the model in Sec. 2.1, and add a CLS token and a 3-layer MLP to the HQ Decoder to serve as the text decoder.

To train the entire model, we follow a three-step process as follows:

1. Pre-training the segmentation model: the backbone image encoder and the HQ decoder are fine-tuned, using the sample data comprising input $\{I, M^{cls}, p^{cls}, t^{cls}\}$ and supervision from M^{GT}.
2. Training the TPP: with the backbone encoders frozen, the text decoder is trained using the sample data input $\{I, M^{cls}, p^{cls}\}$ and supervision from t^{cls}.
3. Training the PSM: with the backbone encoders frozen, the PSM is trained and HQ decoder is further fine-tuned, using the synthetic sequence data

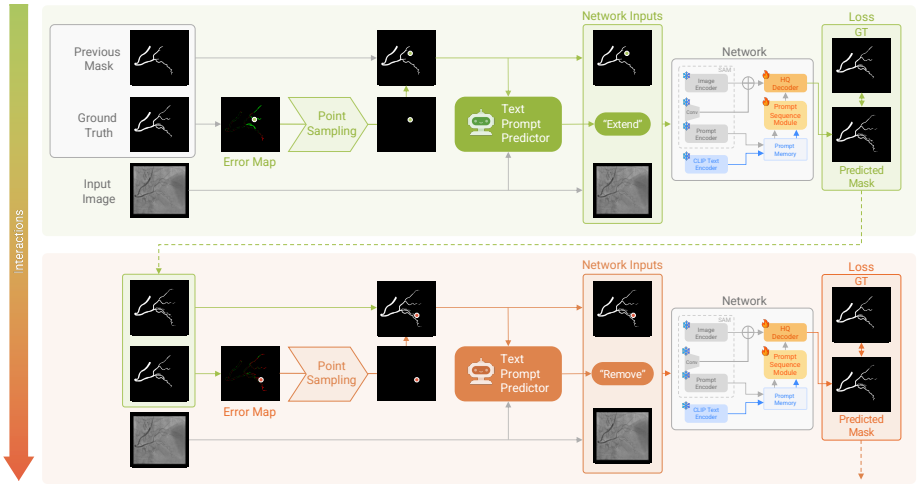

Fig. 3. Training process using generated interactive prompt sequence data. Using the generated text and point prompts from the text prompt predictor and point sampling process, the prompt sequence module is trained and the HQ-decoder is fine-tuned.

with input $\mathcal{D} = \{I, M^k, p^k, t^k\}, k = 0, ..., K-1$ and supervision from M^{GT}. At each step k, M^k is the output of the HQ decoder at step $k-1$, p^k is the center point of the largest connected component of the difference mask between M^k and M^{GT} (prompt point sampling), and t^k is the output of the TPP. Fig. 3 provides a visual summary of this process.

3 Experiments

3.1 Dataset

We conducted experiments using six medical imaging datasets to evaluate the performance of the proposed network. These datasets cover various vascular imaging domains, including retinal fundus images, OCTA (Optical Coherence Tomography Angiography), and X-ray coronary angiography (XCA), and were used to assess the segmentation performance of complex vascular structures.

- FIREFLY [6]: Emphasis is placed on fine vessels localized by aligning fluorescein angiography (FA) and fundus images. Partitioned into 404 training and 45 validation images.
- HRF [1]: High-resolution retinal fundus images, containing both healthy retinas and cases of diabetic retinopathy and glaucoma. Partitioned into 36 training and 9 validation images.
- CHASE [5]: Retinal images from multiethnic children, offering high-resolution vascular segmentation masks. Partitioned into 22 training and 6 validation images.

- OCTA-3mm and OCTA-6mm [13]: Optical Coherence Tomography Angiography images sampled from the OCTA-500 dataset, categorized by the field of view (FOV) captured. OCTA-3mm and OCTA-6mm are partitioned into 140, 10, 50 and 240, 10, 50 training, validation, and test images, respectively.
- ARCADE [17]: X-ray coronary angiography (XCA) images, intended for the segmentation of major coronary arteries. Partitioned into 1000 training and 200 validation images.

3.2 Implementation details

Experiments were conducted on a single RTX 3090 GPU using SAM backbone models from the official repository. The number of training epochs was adjusted per dataset based on the characteristics of each training phase. In all training phases, the batch size was set to 1. The Adam optimizer [10] was used with an initial learning rate of 1e-5. A learning rate scheduler reduced the learning rate by a factor of 0.5 every 10 steps. The loss function varied by phase: Binary Cross Entropy (BCE) loss was used in Phases 1 and 3 for segmentation mask prediction, while Cross Entropy (CE) loss was used in Phase 2 for text prediction. In our experiments, the initial segmentation mask M^0 is generated using nnUNet [8], trained separately for each dataset. To focus on the regions with the largest segmentation errors, images were cropped to 256×256 before applying the interactive segmentation process.

3.3 Evaluation results

Comparative experiments were conducted with SAM [11] and existing single-modal prompt-based interactive models: RITM [21], FocalClick [3], and SimpleClick [15]. Mean Intersection-over-Union (mIoU) and centerline Dice score (clDice) were used as the evaluation metrics. Measurements were made after 10 prompt interactions, using automatic prompts as described in Sec. 2.2 for consistency. All interactive methods, including the proposed one, were either fine-tuned or trained from scratch on each dataset, as appropriate.

Table 1 presents the quantitative evaluations across six datasets, including p-values computed from paired t-tests. The proposed method outperforms the comparison methods on average and achieves higher performance on all datasets except for SimpleClick on the ARCADE dataset. For most comparisons, the p-values indicate rejection of the null hypothesis; however, some higher p-values were observed, likely due to the limited number of images in the datasets. The comparison between our method with and without the PSM shows that most performance gains come from the use of multimodal prompts, with modest additional improvement from learning the prompt sequence.

An important observation is that the proposed method consistently improves upon the baseline nnUNet [8] segmentation, across all datasets. In contrast, other methods often perform worse than the baseline even after 10 prompt interactions. These results highlight the limitations of existing methods in handling complex vascular structures.

Table 1. Quantitative comparative evaluation on various vascular segmentation datasets. mIoU and clDice after 10 prompt interactions are presented, along with p-values computed from paired t-tests.

Model	mIoU@10						
	FIREFLY	HRF	CHASE	OCTA 6mm	OCTA 3mm	ARCADE	avg
nnUNet (baseline) [8]	55.28 (p<0.05)	51.76 (p<0.05)	69.94 (p=0.09)	76.43 (p<0.05)	82.90 (p<0.05)	72.87 (p<0.05)	68.19
SAM [11]	55.62 (p<0.05)	57.52 (p<0.05)	55.49 (p=0.13)	76.35 (p<0.05)	81.77 (p<0.05)	67.32 (p<0.05)	65.67
RITM [21]	35.76 (p<0.05)	11.58 (p<0.05)	0.00 (p<0.05)	61.28 (p<0.05)	62.16 (p<0.05)	81.68 (p=0.13)	42.07
FocalClick [3]	55.24 (p=0.19)	55.97 (p=0.10)	62.21 (p=0.18)	56.77 (p<0.05)	58.25 (p<0.05)	81.13 (p=0.27)	61.59
SimpleClick [15]	51.83 (p<0.05)	56.63 (p<0.05)	50.21 (p<0.05)	70.21 (p<0.05)	72.17 (p<0.05)	**84.55** (p<0.05)	64.26
Ours (w/o PSM)	**57.42** (p=0.64)	**63.38** (p=0.72)	75.79 (p=0.86)	**78.97** (p=0.34)	84.46 (p=0.18)	81.68 (p<0.05)	73.61
Ours (w PSM)	57.26	63.09	**75.92**	78.85	**84.58**	82.43	**73.68**
Model	clDice@10						avg
nnUNet (baseline) [8]	62.12 (p<0.05)	72.61 (p=0.13)	83.99 (p=0.19)	89.64 (p<0.05)	93.05 (p<0.05)	85.17 (p<0.05)	81.09
SAM [11]	63.34 (p=0.06)	73.24 (p=0.10)	65.58 (p=0.12)	88.54 (p<0.05)	92.68 (p<0.05)	79.81 (p<0.05)	77.19
RITM [21]	40.82 (p<0.05)	19.83 (p<0.05)	0.00 (p<0.05)	78.63 (p<0.05)	81.65 (p<0.05)	95.69 (p=0.50)	52.77
FocalClick [3]	61.43 (p=0.05)	74.30 (p=0.33)	77.30 (p=0.18)	71.21 (p<0.05)	70.78 (p<0.05)	92.03 (p<0.05)	74.50
SimpleClick [15]	58.57 (p<0.05)	73.67 (p=0.14)	62.51 (p<0.05)	85.41 (p<0.05)	88.83 (p<0.05)	**97.32** (p<0.05)	77.71
Ours (w/o PSM)	**64.52** (p=0.24)	**77.23** (p=0.73)	**87.54** (p=0.33)	91.14 (p=0.94)	94.14 (p=0.05)	95.20 (p<0.05)	**84.96**
Ours (w PSM)	64.18	76.93	87.06	**91.15**	**94.28**	96.04	84.94

We provide qualitative comparisons for sample images from the HRF and OCTA-6mm datasets in Figures 4 and 5, respectively. While SAM and SimpleClick struggle to differentiate major vessels in the presence of noise, our model maintains robust performance even in noisy environments.

3.4 Ablation Study

We present ablative comparisons for the following PSM configurations in Table 2: the set of encoded point and text prompts are all directly concatenated and fed into the HQ decoder (w/o PSM); the encoded point and text prompts at each step are fed into an LSTM (LSTM); the set of encoded point prompts and text prompts are separately concatenated, and fed into a cross-attention layer [22] as the query and key vectors, respectively (Cross-attn); the set of encoded point and text prompts are all directly concatenated and fed into a self-attention layer (Self-attn, all text); and the set of encoded point prompts and only the encoded

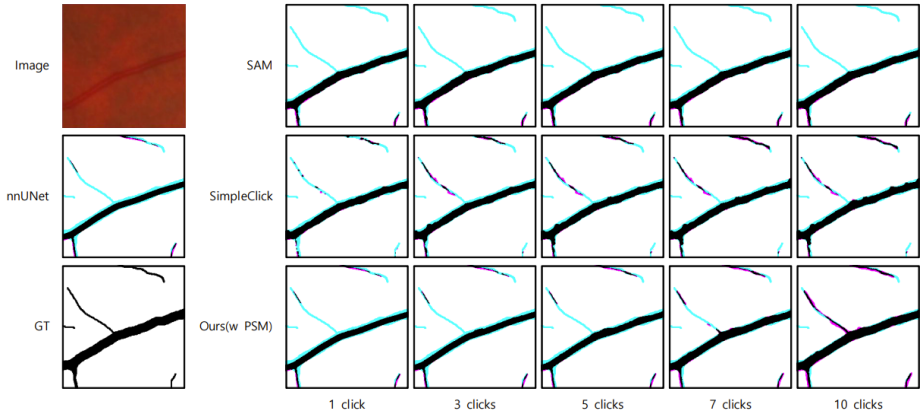

Fig. 4. Qualitative comparison on a sample image from the HRF [1] dataset. Black, cyan, and magenta pixels denote true-positive, false-negative, and false-positive pixels, respectively.

Table 2. Ablative comparative evaluation for various configurations of the PSM. mIoU after 10 prompt interactions (mIoU@10) are presented.

Model	mIoU@10						
	FIREFLY	HRF	CHASE	OCTA 6mm	OCTA 3mm	ARCADE	avg
w/o PSM	57.42	63.38	75.79	78.97	84.46	81.68	73.61
LSTM	57.10	61.80	75.30	**79.11**	84.41	80.92	73.10
Cross-Attn.	**57.51**	**63.59**	75.69	78.82	84.31	81.40	73.55
Self-Attn. (All Text)	57.46	63.01	75.78	78.58	84.39	81.74	73.49
Self-Attn. (Last Text)	57.26	63.09	**75.92**	78.85	**84.58**	**82.43**	**73.68**

last text prompt are directly concatenated and fed into a self-attention layer (Self-attn, last text).

These results highlight the different properties of the point and text prompts. That is, while PSM configurations using the whole set of text prompts actually degrade the performance relative to simple concatenation. But when using only the last text prompt together with all point prompts, the performance is improved. Thus, this PSM configuration was used in Table 1.

4 Conclusion

We present a novel interactive framework that learns sequences of multimodal prompts for segmentation of complex vascular structures. The proposed method acts to avoid critical limitations, where interactions actually degrade baseline segmentation results, and enable user intended improvements of segmentation masks. We hope this work is an initial step in the development of more intelligent

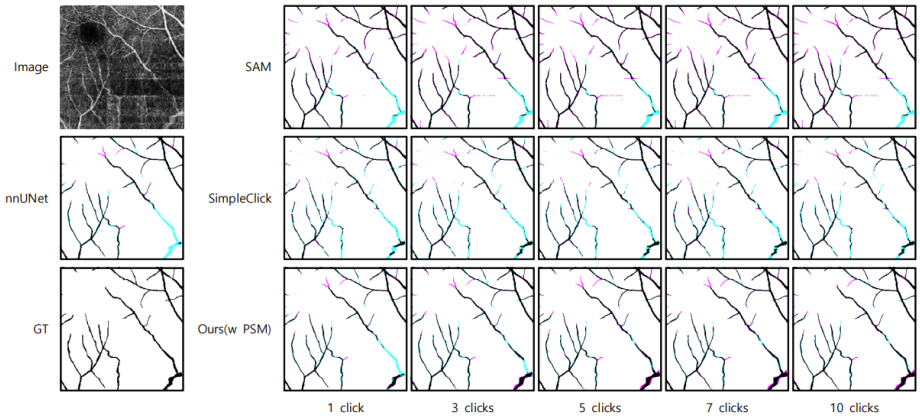

Fig. 5. Qualitative comparison on a sample image from the OCTA-6mm [13] dataset. Black, cyan, and magenta pixels denote true-positive, false-negative, and false-positive pixels, respectively.

interactive tools that can fully realize the operators' intented target mask with minimal input.

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

References

1. Budai, A., Bock, R., Maier, A., Hornegger, J., Michelson, G.: Robust vessel segmentation in fundus images. International journal of biomedical imaging **2013**(1), 154860 (2013)
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
3. Chen, X., Zhao, Z., Zhang, Y., Duan, M., Qi, D., Zhao, H.: Focalclick: Towards practical interactive image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1300–1309 (2022)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Fraz, M.M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A.R., Owen, C.G., Barman, S.A.: An ensemble classification-based approach applied to retinal blood vessel segmentation. IEEE Transactions on Biomedical Engineering **59**(9), 2538–2548 (2012)
6. Go, S., Kim, J., Noh, K.J., Park, S.J., Lee, S.: Combined deep learning of fundus images and fluorescein angiography for retinal artery/vein classification. IEEE Access **10**, 70688–70698 (2022)

7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
8. 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)
9. Ke, L., Ye, M., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F., et al.: Segment anything in high quality. Advances in Neural Information Processing Systems **36**, 29914–29934 (2023)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
12. Lee, T.C., Kashyap, R.L., Chu, C.N.: Building skeleton models via 3-d medial surface axis thinning algorithms. CVGIP: graphical models and image processing **56**(6), 462–478 (1994)
13. Li, M., Huang, K., Xu, Q., Yang, J., Zhang, Y., Ji, Z., Xie, K., Yuan, S., Liu, Q., Chen, Q.: Octa-500: a retinal dataset for optical coherence tomography angiography study. Medical image analysis **93**, 103092 (2024)
14. Liu, Q., Cho, J., Bansal, M., Niethammer, M.: Rethinking interactive image segmentation with low latency high quality and diverse prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3773–3782 (2024)
15. Liu, Q., Xu, Z., Bertasius, G., Niethammer, M.: Simpleclick: Interactive image segmentation with simple vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22290–22300 (2023)
16. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
17. Popov, M., Amanturdieva, A., Zhaksylyk, N., Alkanov, A., Saniyazbekov, A., Aimyshev, T., Ismailov, E., Bulegenov, A., Kuzhukeyev, A., Kulanbayeva, A., et al.: Dataset for automatic region-based coronary artery disease diagnostics using x-ray angiography images. Scientific data **11**(1), 20 (2024)
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. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
20. Serra, J.: Image Analysis and Mathematical Morphology. Academic Press, Inc., USA (1983)
21. Sofiiuk, K., Petrov, I.A., Konushin, A.: Reviving iterative training with mask guidance for interactive segmentation. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 3141–3145. IEEE (2022)
22. 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)
23. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021)