Spatial-Temporal Memory Filtering SAM for Lesion Segmentation in Breast Ultrasound Videos

Zhengzheng Tu[1], Liang Zong[1], Bo Jiang[1] (✉), Haowen Wang[1] (✉), Kunpeng Wang[1], and Chaoxue Zhang[2]

[1] School of Computer Science and Technology, Anhui University, Hefei, China
[2] Department of Ultrasound, the First Affiliated Hospital of Anhui Medical University, Hefei, 230022, China
zhengzhengahu@163.com, jiangbo@ahu.edu.cn, wanghaowen@ahu.edu.cn

Abstract. Lesion segmentation in breast ultrasound videos plays a crucial role in the early detection and intervention of breast cancer. However, it remains a challenging task due to blurred lesion boundaries, substantial background noise, and significant scale variations of lesions across frames. Existing methods typically rely on selecting preceding frames for rudimentary temporal integration but fail to achieve satisfactory segmentation performance. In this paper, we propose STMFSAM, a novel Spatio-Temporal Memory Filtering SAM network, designed to leverage the powerful feature representation and modeling capabilities of SAM for lesion segmentation in breast ultrasound videos. Specifically, we introduce a memory mechanism that stores and propagates essential spatio-temporal features across frames. To enhance segmentation accuracy, we select three relevant reference frames from the memory bank as dense prompts for SAM, enabling it to retain long-term contextual information and effectively guide the segmentation of subsequent frames. To further mitigate the impact of background noise, we present the Spatio-Temporal Memory Filtering module, which selectively refines the memory content by filtering out irrelevant or noisy information. This ensures that only meaningful and informative features are retained for segmentation. We conduct extensive experiments on the UVBSL200 breast ultrasound video dataset, demonstrating that STMFSAM outperforms existing methods. Additionally, to highlight our model's generalization capability, we achieve competitive results on two video polyp segmentation datasets. The code is available at https://github.com/tzz-ahu/STMFSAM.

Keywords: Breast lesion · Ultrasound video · Segmentation · SAM .

1 Introduction

Breast cancer stands as the most prevalent cancer in females worldwide, with recent statistics indicating that one in every eight newly diagnosed cancer cases is attributed to breast cancer [1]. Given its complex etiology and individual variability, early screening is crucial for timely intervention. Due to noninvasive, cost-effective, and real-time, ultrasound imaging has become the most widely used method in breast cancer screening [22]. However, ultrasound image interpretation relies heavily on clinician expertise, leading to significant variability in assessment results, even by the same operator at different times [21]. This raises concerns about diagnostic consistency and accuracy. To alleviate clinicians' workload and enhance diagnostic precision, automated segmentation techniques are employed to provide lesion morphology information. These techniques assist physicians in evaluating benignity, malignancy, aggressiveness, and staging. Hence, lesion segmentation in breast ultrasound videos has notable clinical value and broad application potential.

Existing video-based methods [6,11] focus mainly on short-term temporal information between adjacent frames. For example, FLA-Net [6] uses a contrastive loss to reduce discrepancies in lesion locations, while TMFF [11] relies on previous frame segmentation results as prior information for stability. However, these methods struggle to capture temporal evolution, especially in tracking lesion shape changes and contours. Additionally, they fail to obtain global information because of the CNN backbones' limited receptive field, making it hard to distinguish background noise from foreground targets, which impacts segmentation accuracy.

To cope with these issues, we propose STMFSAM, a novel Spatio-Temporal Memory Filtering framework for lesion segmentation in breast ultrasound videos. Given SAM's [12] ability to learn discriminative features and model global context, we employ SAM as the backbone of our network. To handle lesion shape variations and blurred contours in ultrasound videos, we introduce a memory mechanism to store and propagate critical spatio-temporal features across frames. From this memory bank, we select three reference frames—the first frame of the video, the previous frame and the most similar memory frame to the current one—to provide spatial and positional context. These frames serve as dense prompts, allowing SAM to retain long-term context and guide segmentation in subsequent frames. Additionally, we develop a Spatio-Temporal Memory Filtering module to optimize the memory bank content, removing irrelevant features and preserving the most useful information for accurate segmentation. This filtering improves segmentation accuracy and suppresses background noise. We assess the performance of STMFSAM on the UVBSL200 breast ultrasound dataset [11], achieving state-of-the-art results compared to existing methods and showcasing robust generalization on two video polyp segmentation datasets. Our key contributions are as follows:

(1)We introduce a novel breast ultrasound video segmentation model incorporating SAM, which improves segmentation accuracy by utilizing spatio-temporal information from the memory bank.

(2)We design a spatio-temporal memory filtering module to enhance memory features, reducing redundant information and noise.

(3)We validate our approach on a large-scale breast ultrasound dataset, achieving outstanding performance and demonstrating strong generalization capabilities in video polyp segmentation tasks.

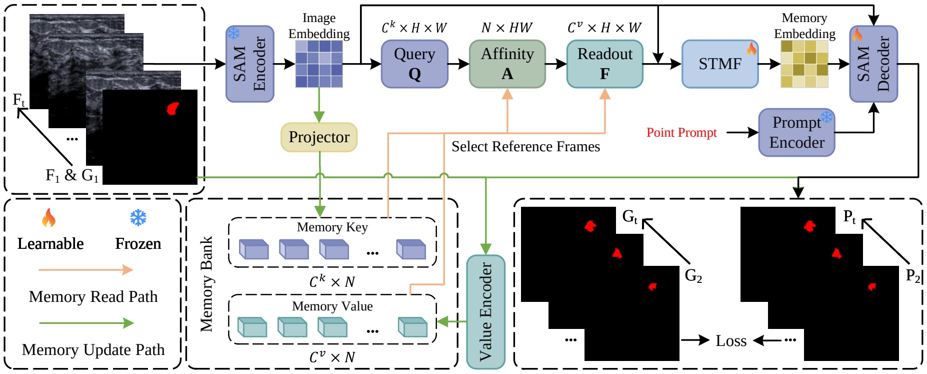

Fig. 1. Overview of STMFSAM, which consists of SAM, Memory Mechanism, and Spatial-Temporal Memory Filtering Module.

2 Method

Overview. The structure of our proposed Spatio-Temporal Memory Filtering SAM (STMFSAM) is illustrated in Figure 1. It comprises three main components: the SAM, memory mechanism, and filtering module. The SAM encoder transforms the input image into embeddings, while the prompt encoder encodes foreground points as sparse prompts. The memory mechanism includes two key operations: storing and reading. For storage, image embeddings are projected into memory space as keys, while frames and their corresponding masks are encoded into memory values by a value encoder. During memory reading, we compute the current frame and stored frames' affinity, and the most relevant memory values are selected as reference features. These features are then processed by the filtering module to reduce background noise and redundant information, providing refined dense prompts to the SAM decoder. The decoder integrates all the prompts and features to predict the final masks.

2.1 Memory Mechanism

The memory mechanism, as shown in Figure 1, perform the storage of memory keys and values and their retrieval on readout, a process that occurs in real time as each frame is processed.

Memory storing. The inputs of our method are t video frames $F_1, F_2, ..., F_t \in \mathbb{R}^{C\times H\times W}$ and groundtruth mask $G_1 \in \mathbb{R}^{C\times H\times W}$ corresponding to F_1, where C, H and W are the channel, height and width of video frames or the groundtruth masks and t denotes the index of the frame in the sequence. This follows a common paradigm in semi-supervised video object segmentation, where the initial frame's annotation is provided to initialize and guide the segmentation process for the entire sequence. For each frame F_i, after being processed by the SAM encoder E_s, an image embedding I_i is generated. This embedding is then projected into the memory space through a projection layer, forming the memory key $K_i \in \mathbb{R}^{C^k \times H \times W}$. When groundtruth (with the first frame) or the final predicted mask is available, the current frame F_i and the mask P_i (where $P_1 = G_1$ for $i = 1$) are encoded by using a dedicated value encoder E_v into the memory value $V_i \in \mathbb{R}^{C^v \times H \times W}$. Here, C^k and C^v denotes the dimension of key and value in memory space, respectively. The process can be formulated as:

$$K_i = Proj(E_s(F_i)) \tag{1}$$

$$V_i = E_v(F_i, P_i), (i = 1, P_1 = G_1) \tag{2}$$

where $1 \leq i \leq t$. The keys and values are concatenated along a specific dimension in the memory bank, and then we can search for the matching memory values based on the indices of the memory keys.This design ensures that the memory bank not only stores the spatial features of individual frames but also captures the temporal dynamics by linking the information across consecutive frames.

Memory Reading. The memory reading operation depicted in Figure 1 elucidates the process of retrieving spatio-temporal information from the memory bank to assist in generating the segmentation result for the current frame. Specifically, following the memory storage phase, our model maintains a memory bank that stores key-value pairs from previous frames. Our objective is to obtain the readout features $R_i \in \mathbb{R}^{C^v \times H \times W}$, which are derived from the memory values and affinity matrices, to be utilized as dense prompts.

Depending on the varying number of stored memory frames within the memory bank, distinct memory retrieval methods are employed. In scenarios where the number of memory frames is less than or equal to three, we directly use all memory frames in the memory bank as reference frames. Conversely, when the number of memory frames exceeds three, we utilize three memory frames: (1) the first frame I_1, which contains the ground truth mask, (2) the most recent frame I_{i-1}, and (3) a memory frame I_m selected based on the highest similarity between the current frame I_i and any prior frame $I_2, I_3, ..., I_{i-2}$. For the sake of clarity, we assume that the current frame query is denoted as Q_i, the reference memory key as K_{i-1}, and the reference memory value as V_{i-1}. The overall process can be formulated as:

$$A(K_{i-1}, Q_i) = softmax(S(K_{i-1}, Q_i)) \tag{3}$$

$$R_i = V_{i-1} \cdot A(K_{i-1}, Q_i) \tag{4}$$

where S denotes the anisotropic L2 similarity [4] function. This multi-faceted approach ensures that the segmentation process leverages both temporal continuity (recent frames) and long-term information (ground truth and similar frames), improving the accuracy of segmentation across different frames.

2.2 Spatial-Temporal Memory Filtering

Ultrasound images, with their inherent complex noise, can lead to noisy feature representations encoded by the encoder. If directly used for segmentation, these

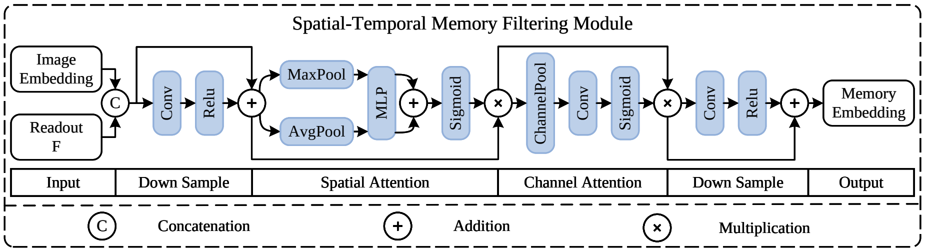

Fig. 2. Details of STMF Module, which leverages downsampling, spatial attention, and channel attention mechanism to filter and refine the features read from the memory bank.

representations may result in poor foreground-background separation. To mitigate this, we propose a Spatio-Temporal Memory Filtering (STMF) module that enhances target localization and suppresses background noise.

First, the image embedding I_i and the readout feature R_i are concatenated along the channel dimension to form the input F_{in}, combining spatial information with high-level semantic features. This integration allows the module to exploit both low- and high-level features, essential for distinguishing foreground from noise. Downsampling is then applied to reduce the spatial dimensions, which simplifies subsequent processing. Next, spatial attention is applied to emphasize regions containing the foreground target while suppressing background noise:

$$F_s = \sigma(f([AvgP(F_{in}); MaxP(F_{in})])) \tag{5}$$

where σ is the sigmoid activation function, f represents a convolution, and $[;]$ denotes concatenation of average pooling $AvgP$ and max pooling $MaxP$ results along the channel axis. Channel attention further refines feature representations by amplifying relevant channels:

$$F_c = \sigma(MLP(AvgP(F_s)) + MLP(MaxP(F_s))) \tag{6}$$

This operation reduces noise by prioritizing informative channels and diminishing less relevant ones. A second downsampling step ensures that only the most salient features remain, minimizing noise impact and producing compact memory embeddings. Residual connections are incorporated throughout to support information and gradient flow, enhancing model robustness.

3 Experiment

Implementation Details. For the UVBLS200 dataset, we adhere to the official data split of 180 videos for training and 20 for testing. All experiments were conducted using a fixed random seed to ensure reproducibility. The SAM and prompt encoders are initialized using weights pre-trained on ImageNet, whereas the remaining components are trained from scratch. Video frames are resized to

Table 1. Segmentation performance comparison between the proposed method and state-of-the-art approaches on the UVBLS200 dataset.

Input	**Method**	**Year**	**Dice ↑**	**IoU ↑**	**Recall ↑**	**MAE ↓**
Image	UNet++ [13]	2018	0.723	0.576	0.664	0.054
	HarDNet [14]	2019	0.823	0.727	0.823	0.035
	MSNet [15]	2021	0.800	0.700	0.769	0.041
	TRUNet [16]	2022	0.819	0.724	0.828	0.038
	UCTNet [17]	2022	0.825	0.721	0.846	0.037
	SAM [12]	2023	0.631	0.514	0.683	0.179
	SAMUS [18]	2023	0.838	0.755	0.872	0.044
Video	STM [19]	2019	0.821	0.729	0.857	0.039
	AFB-URR [20]	2020	0.811	0.713	0.794	0.037
	STCN [3]	2021	0.834	0.742	0.845	0.033
	DCFNet [23]	2021	0.804	0.707	0.794	0.035
	XMem [4]	2022	0.851	0.762	0.861	0.026
	UFO [24]	2023	0.789	0.680	0.813	0.040
	TMFF [11]	2024	0.841	0.752	0.888	0.035
	Ours	2024	**0.872**	**0.787**	**0.897**	**0.022**

256x256 and undergo data augmentation techniques such as random cropping, flipping, and affine transformations. The model is built using PyTorch and optimized with AdamW [25], employing a learning rate of 1e-5, weight decay of 0.05, and a linear warmup phase spanning 250 iterations, followed by step-wise decay. To enhance training efficiency and minimize memory consumption, automatic mixed precision (AMP) is utilized. The memory key and value dimensions are configured to 64 and 512, respectively, with a memory bank capacity of 8. ResNet18 is used for value encoding. We set batch size to 1 and use a single NVIDIA RTX 3090 GPU to conduct the training process. We acknowledge that the introduction of the memory mechanism and the STMF module adds a moderate computational overhead to the baseline SAM architecture. However, we posit that this trade-off is justified by the substantial gains in segmentation accuracy and robustness, which are critical for addressing the inherent challenges of ultrasound video analysis.

3.1 Comparisons with State-of-the-arts

To validate the effectiveness of our approach, we conducted a quantitative comparison with several state-of-the-art (SOTA) methods on the UVBLS200 dataset. These methods include seven image-based methods: UNet++ [13], HarDNet [14], MSNet [15], TRUNet [16], UCTNet [17], SAM [12], SAMUS [18], and seven video-based methods: STM [19], AFB [20], STCN [3], DCF-Net [23], ,XMem [4], UFO [24], TMFF [11]. To ensure a balanced and impartial comparison, we obtain the segmentation results of the fifteen approaches either by utilizing their publicly accessible implementations or by developing our own versions.

Quantitative Comparisons. Table 1 presents a performance comparison between our model and other existing methods. SAM struggles to learn effective

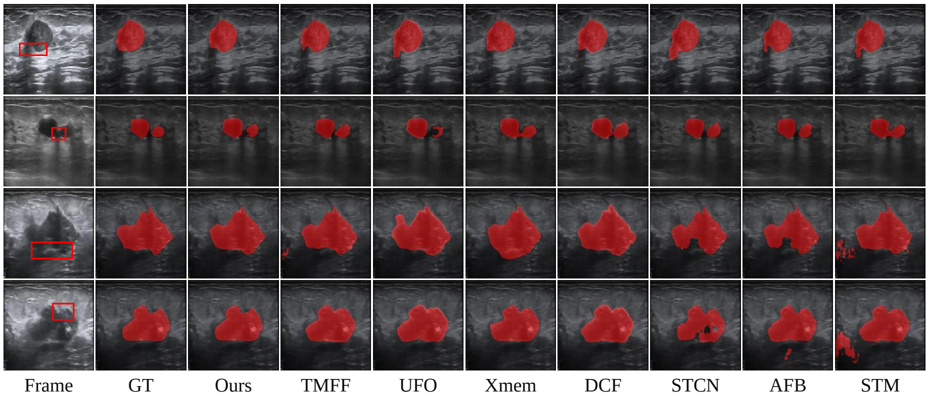

Fig. 3. Visual comparison with state-of-the-art methods on the UVBLS200 test set. The leftmost column represents the original frames, with the most challenging aspects of lesion segmentation highlighted by red boxes. In the remaining columns, the red regions overlaid in each image represent the groundtruth or prediction of the breast lesion.

medical feature representations due to data scale limitations, leading to suboptimal segmentation results. SAMUS, designed for ultrasound images, performs better but is still outperformed by our model across all evaluation metrics. Notably, for video sequences, our method achieves a 3.1% and 3.5% improvement in Dice and IoU scores, respectively, compared to the benchmark method TMFF. This highlights our model's ability to effectively capture spatio-temporal information and remain robust against motion artifacts and noise in ultrasound imaging. The improvements are attributed to the spatio-temporal information fusion and filtering module integrated into our model.

Qualitative Comparisons. Figure 3 provides a visual comparison of segmentation outcomes generated by our model and leading state-of-the-art approaches. Our method demonstrates superior accuracy in segmenting breast lesions, effectively handling challenges such as varying lesion sizes, irregular shapes, and inconsistent intensity across frames. In contrast, other methods often over-segment or under-segment boundaries, particularly in cases of blurred edges and high background noise typical of ultrasound images.

3.2 Ablation Study

We performed ablation experiments to assess the impact of our model's key components: the Spatio-Temporal Memory Filtering Module (STMF), point-based prompting, and the Reference Frame Selection Algorithm (RFSA). The results are presented in Table 2. First, removing the STMF module resulted in a consistent performance drop across all metrics. While the quantitative decrease in Dice and IoU is modest, this finding is consistent with the module's designated function: to refine retrieved memory features by suppressing background noise and

Table 2. Ablation analysis of STMFSAM's components on the UVBLS200 dataset.

Setting	Dice ↑	IoU ↑	Recall ↑	MAE ↓
full model	**0.872**	**0.787**	**0.897**	**0.022**
w/o STMF	0.866	0.778	0.884	0.023
w/o point prompt	0.863	0.774	0.879	0.025
w/o RFSA	0.856	0.748	0.864	0.033

irrelevant information. This filtering process is crucial for achieving stable and qualitatively superior segmentations, especially in frames with high noise levels, even if its impact on aggregate metrics is not drastic. Second, eliminating point-based prompting caused a slight but noticeable performance decline, confirming the value of providing explicit spatial guidance to the SAM decoder. Finally, replacing our RFSA with a random frame selection strategy led to a substantial performance deterioration. This clearly underscores the effectiveness of our selection strategy—leveraging the first, previous, and most similar frames—in capturing critical spatio-temporal dependencies for accurate video segmentation.

3.3 Generalization Capability

To assess the generalization capability of our STMFSAM, we extended its evaluation to the task of video polyp segmentation (VPS). Adhering to the experimental framework described in a recent study on video polyp segmentation [5], we retrained our network and evaluated its performance on two benchmark datasets: CVC-300-TV and CVC-612-V. The quantitative results, compared against state-of-the-art methods, are presented in Table 3. We employed several evaluation metrics, including Dice coefficient (Dice), Enhanced-alignment measure (E_ϕ) [27] , Intersection over Union (IoU), Mean Absolute Error (MAE) and S-measure (S_α) [26]. Our approach consistently surpasses existing methods across all metrics, clearly demonstrating its superior capability in accurately segmenting polyp regions.

Table 3. Performance evaluation results across two video polyp segmentation datasets.

	Metrics	ACSNet [2]	PraNet [9]	PNSNet [5]	FLA-Net [6]	STCN [3]	XMem [4]	Ours
CVC-300-TV	Dice ↑	0.738	0.739	0.840	0.874	0.867	0.893	**0.911**
	IoU ↑	0.632	0.645	0.745	0.789	0.784	0.826	**0.842**
	S_α ↑	0.837	0.833	0.909	0.907	0.896	0.915	**0.947**
	E_ϕ ↑	0.871	0.852	0.921	0.969	0.962	0.967	**0.981**
	MAE ↓	0.016	0.016	0.013	0.010	0.014	0.009	**0.005**
CVC-612-V	Dice ↑	0.804	0.869	0.873	0.885	0.876	0.889	**0.905**
	IoU ↑	0.712	0.799	0.800	0.814	0.811	0.818	**0.846**
	S_α ↑	0.847	0.915	0.923	0.920	0.917	0.927	**0.938**
	E_ϕ ↑	0.887	0.936	0.944	0.963	0.942	0.956	**0.979**
	MAE ↓	0.054	0.013	0.012	0.012	0.013	0.010	**0.007**

4 Conclusion

In this work, we enhance the Segment Anything Model by incorporating a memory mechanism design, enabling SAM to effectively leverage temporal information from video sequences. This modification extends SAM's applicability to lesion segmentation in breast ultrasound videos. Furthermore, to address the distinct challenges of ultrasonic imaging, we introduce a Spatio-Temporal Memory Filtering module aimed at reducing the impact of noise on the model's learning process. Extensive experiments on the UVBLS200 and VPS datasets demonstrate that our methodology not only achieves state-of-the-art performance, but also exhibits commendable generalizability. Looking ahead, we intend to further investigate the potential of the SAM model in the realm of medical video segmentation to address the demands of clinical applications.

Acknowledgments. This work was supported by National Natural Science Foundation of China under Grant 62376005; Natural Science Foundation of Anhui Province under Grant 2408085J037.

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

References

1. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)
2. Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., Yu, Y.: Adaptive context selection for polyp segmentation. In: MICCAI 2020, pp. 253–262. Springer (2020)
3. Cheng, H.K., Tai, Y.W., Tang, C.K.: Rethinking space-time networks with improved memory coverage for efficient video object segmentation. Advances in Neural Information Processing Systems **34**, 11781–11794 (2021)
4. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: European Conference on Computer Vision, pp. 640–658. Springer (2022)
5. Ji, G.P., Chou, Y.C., Fan, D.P., Chen, G., Fu, H., Jha, D., Shao, L.: Progressively normalized self-attention network for video polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 142–152. Springer (2021)
6. Lin, J., Dai, Q., Zhu, L., Fu, H., Wang, Q., Li, W., Rao, W., Huang, X., Wang, L.: Shifting more attention to breast lesion segmentation in ultrasound videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 497–507. Springer (2023)
7. Zhuang, Z., Li, N., Joseph Raj, A.N., Mahesh, V.G.V., Qiu, S.: An RDAU-NET model for lesion segmentation in breast ultrasound images. PloS one **14**(8), e0221535 (2019)

8. Zhao, Y., Que, D., Tan, J., Xiao, Y., Yu, Y.: Automated breast lesion segmentation from ultrasound images based on ppu-net. In: 2019 International Conference on Medical Imaging Physics and Engineering (ICMIPE), pp. 1–4. IEEE (2019)
9. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 263–273. Springer (2020)
10. Byra, M., Jarosik, P., Szubert, A., Galperin, M., Ojeda-Fournier, H., Olson, L., O'Boyle, M., Comstock, C., Andre, M.: Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network. Biomedical Signal Processing and Control **61**, 102027 (2020)
11. Tu, Z., Zhu, Z., Duan, Y., Jiang, B., Wang, Q., Zhang, C.: A Spatial-Temporal Progressive Fusion Network for Breast Lesion Segmentation in Ultrasound Videos. arXiv preprint arXiv:2403.11699 (2024)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P., Girshick, R.: Segment Anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015–4026 (2023)
13. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: DLMIA 2018, ML-CDS 2018, pp. 3–11. Springer (2018)
14. Chao, P., Kao, C.Y., Ruan, Y.S., Huang, C.H., Lin, Y.L.: Hardnet: A low memory traffic network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3552–3561 (2019)
15. Zhao, X., Zhang, L., Lu, H.: Automatic polyp segmentation via multi-scale subtraction network. In: MICCAI 2021, pp. 120–130. Springer (2021)
16. Tomar, N.K., Shergill, A., Rieders, B., Bagci, U., Jha, D.: TransResU-Net: Transformer based ResU-Net for real-time colonoscopy polyp segmentation. arXiv preprint arXiv:2206.08985 (2022)
17. Wang, H., Cao, P., Wang, J., Zaiane, O.R.: Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 3, pp. 2441–2449 (2022)
18. Lin, X., Xiang, Y., Zhang, L., Yang, X., Yan, Z., Yu, L.: SAMUS: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. arXiv preprint arXiv:2309.06824 (2023)
19. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9226–9235 (2019)
20. Liang, Y., Li, X., Jafari, N., Chen, J.: Video object segmentation with adaptive feature bank and uncertain-region refinement. Advances in Neural Information Processing Systems **33**, 3430–3441 (2020)
21. Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S.X., Ni, D., Wang, T.: Deep learning in medical ultrasound analysis: a review. Engineering **5**(2), 261–275 (2019)
22. Sood, R., Rositch, A. F., Shakoor, D., Ambinder, E., Pool, K.-L., Pollack, E., Mollura, D. J., Mullen, L. A., Harvey, S. C.: Ultrasound for breast cancer detection globally: a systematic review and meta-analysis. Journal of global oncology (2019)
23. Zhang, M., Liu, J., Wang, Y., Piao, Y., Yao, S., Ji, W., Li, J., Lu, H., Luo, Z.: Dynamic context-sensitive filtering network for video salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1553–1563 (2021)

24. Su, Y., Deng, J., Sun, R., Lin, G., Su, H., Wu, Q.: A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. IEEE Transactions on Multimedia **26**, 313–325 (2023)
25. Loshchilov, I.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
26. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4548–4557 (2017)
27. Fan, D.P., Ji, G.P., Qin, X., Cheng, M.M.: Cognitive vision inspired object segmentation metric and loss function. Scientia Sinica Informationis **6**(6), 5 (2021)