Shifting More Attention to Breast Lesion Segmentation in Ultrasound Videos

Junhao Lin[1], Qian Dai[1], Lei Zhu[2,3(✉)], Huazhu Fu[4], Qiong Wang[5], Weibin Li[6], Wenhao Rao[1], Xiaoyang Huang[1(✉)], and Liansheng Wang[1]

[1] School of Informatics, Xiamen University, Xiamen, China
xyhuang@xmu.edu.cn
[2] ROAS Thrust, System Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
leizhu@ust.hk
[3] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong SAR, China
[4] Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore, Singapore
[5] Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[6] School of Medicine, Xiamen University, Xiamen, China

Abstract. Breast lesion segmentation in ultrasound (US) videos is essential for diagnosing and treating axillary lymph node metastasis. However, the lack of a well-established and large-scale ultrasound video dataset with high-quality annotations has posed a persistent challenge for the research community. To overcome this issue, we meticulously curated a US video breast lesion segmentation dataset comprising 572 videos and 34,300 annotated frames, covering a wide range of realistic clinical scenarios. Furthermore, we propose a novel frequency and localization feature aggregation network (FLA-Net) that learns temporal features from the frequency domain and predicts additional lesion location positions to assist with breast lesion segmentation. We also devise a localization-based contrastive loss to reduce the lesion location distance between neighboring video frames within the same video and enlarge the location distances between frames from different ultrasound videos. Our experiments on our annotated dataset and two public video polyp segmentation datasets demonstrate that our proposed FLA-Net achieves state-of-the-art performance in breast lesion segmentation in US videos and video polyp segmentation while significantly reducing time and space complexity. Our model and dataset are available at https://github.com/jhl-Det/FLA-Net.

Keywords: Ultrasound Video · Breast lesion · Segmentation

1   Introduction

Axillary lymph node (ALN) metastasis is a severe complication of cancer that can have devastating consequences, including significant morbidity and mortality. Early detection and timely treatment are crucial for improving outcomes and reducing the risk of recurrence. In breast cancer diagnosis, accurately segmenting breast lesions in ultrasound (US) videos is an essential step for computer-aided diagnosis systems, as well as breast cancer diagnosis and treatment. However, this task is challenging due to several factors, including blurry lesion boundaries, inhomogeneous distributions, diverse motion patterns, and dynamic changes in lesion sizes over time [12].

Table 1. Statistics of existing breast lesion US videos datasets and the proposed dataset. #**videos**: numbers of videos. #**AD**: number of annotated frames. **BBox**: whether provide bounding box annotation. **BBox**: whether provide segmentation mask annotation. **BM**: whether provide lesion classification label (Benign or Malignant). **PA**: whether provide axillary lymph node (ALN) metastasis label (Presence or Absence).

Dataset	Year	# videos	# AF	BBox	Mask	BM	PA
Li et al. [10]	2022	63	4,619	×	✓	✓	×
Lin et al. [12]	2022	188	25,272	✓	×	✓	×
Ours	2023	**572**	**34,300**	✓	✓	✓	✓

The work presented in [10] proposed the first pixel-wise annotated benchmark dataset for breast lesion segmentation in US videos, but it has some limitations. Although their efforts were commendable, this dataset is private and contains only 63 videos with 4,619 annotated frames. The small dataset size increases the risk of overfitting and limits the generalizability capability. In this work, **we collected a larger-scale US video breast lesion segmentation dataset** with 572 videos and 34,300 annotated frames, of which 222 videos contain ALN metastasis, covering a wide range of realistic clinical scenarios. Please refer to Table 1 for a detailed comparison between our dataset and existing datasets.

Although the existing benchmark method DPSTT [10] has shown promising results for breast lesion segmentation in US videos, it only uses the ultrasound image to read memory for learning temporal features. However, ultrasound images suffer from speckle noise, weak boundaries, and low image quality. Thus, there is still considerable room for improvement in ultrasound video breast lesion segmentation. To address this, **we propose a novel network called Frequency and Localization Feature Aggregation Network (FLA-Net)** to improve breast lesion segmentation in ultrasound videos. Our FLA-Net learns frequency-based temporal features and then uses them to predict auxiliary breast lesion location maps to assist the segmentation of breast lesions in video frames. Additionally, we devise a contrastive loss to enhance the breast lesion location
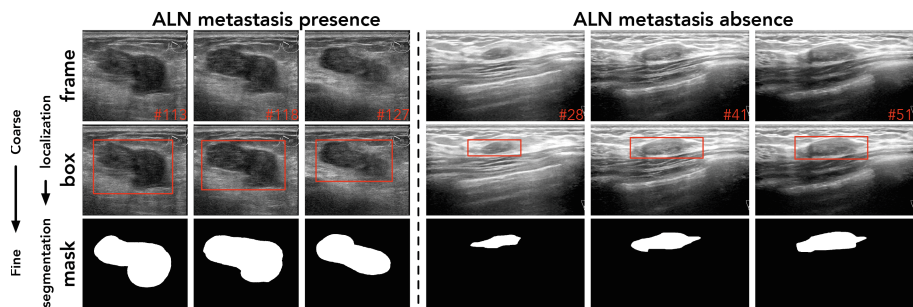
Fig. 1. Examples of our ultrasound video dataset for breast lesion segmentation.

similarity of video frames within the same ultrasound video and to prohibit location similarity of different ultrasound videos. The experimental results unequivocally showcase that our network surpasses state-of-the-art techniques in the realm of both breast lesion segmentation in US videos and two video polyp segmentation benchmark datasets (Fig. 1).

2 Ultrasound Video Breast Lesion Segmentation Dataset

To support advancements in breast lesion segmentation and ALN metastasis prediction, we collected a dataset containing 572 breast lesion ultrasound videos with 34,300 annotated frames. Table 1 summarizes the statistics of existing breast lesion US video datasets. Among 572 videos, 222 videos with ALN metastasis. Nine experienced pathologists were invited to manually annotate breast lesions at each video frame. Unlike previous datasets [10,12], our dataset has a reserved validation set to avoid model overfitting. The entire dataset is partitioned into training, validation, and test sets in a proportion of 4:2:4, yielding a total of 230 training videos, 112 validation videos, and 230 test videos for comprehensive benchmarking purposes. Moreover, apart from the segmentation annotation, our dataset also includes lesion bounding box labels, which enables benchmarking breast lesion detection in ultrasound videos. More dataset statistics are available in the *Supplementary*.

3 Proposed Method

Figure 2 provides a detailed illustration of the proposed frequency and localization feature aggregation network (FLA-Net). When presented with an ultrasound frame denoted as I_t along with its two adjacent video frames (I_{t-1} and I_{t-2}), our initial step involves feeding them through an Encoder, specifically the Res2Net50 architecture [6], to acquire three distinct features labeled as f_t, f_{t-1}, and f_{t-2}. Then, we devise a frequency-based feature aggregation (FFA) module to integrate frequency features of each video frame. After that, we pass the output
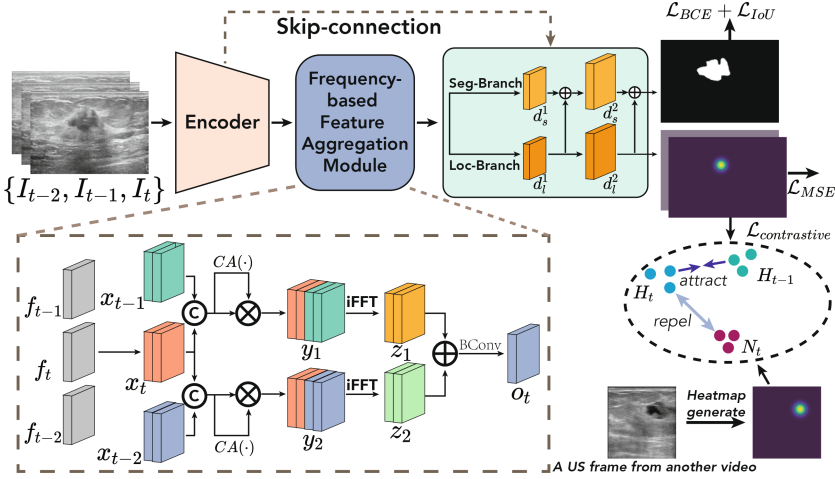
Fig. 2. Overview of our FLA-Net. Our network takes an ultrasound frame I_t and its adjacent two frames (I_{t-1} and I_{t-2}) as input. Three frames are first passed through an encoder to learn three CNN features (f_t, f_{t-1}, and f_{t-2}). Then Frequency-based Feature Aggregation Module is then used to aggregate these features and the aggregated feature map is then passed into our two-branch decoder to predict the breast lesion segmentation mask of I_t, and a lesion localization heatmap. Moreover, we devise a location-aware contrastive loss (see $\mathcal{L}_{contrastive}$) to reduce location distance of frames from the same video and enlarge the location distance of different video frames.

features o_t of the FFA module into two decoder branches (similar to the UNet decoder [14]): one is the localization branch to predict the localization map of the breast lesions, while another segmentation branch integrates the features of the localization branch to fuse localization feature for segmenting breast lesions. Moreover, we devise a location-based contrastive loss to regularize the breast lesion locations of inter-video frames and intra-video frames.

3.1   Frequency-Based Feature Aggregation (FFA) Module

According to the spectral convolution theorem in Fourier theory, any modification made to a single value in the spectral domain has a global impact on all the original input features [1]. This theorem guides the design of FFA module, which has a global receptive field to refine features in the spectral domain. As shown in Fig. 2, our FFA block takes three features ($f_t \in \mathbb{R}^{c \times h \times w}$, $f_{t-1} \in \mathbb{R}^{c \times h \times w}$, and $f_{t-2} \in \mathbb{R}^{c \times h \times w}$) as input. To integrate the three input features and extract relevant information while suppressing irrelevant information, our FFA block first employs a Fast Fourier Transform (FFT) to transform the three input features into the spectral domain, resulting in three corresponding spectral domain features ($\hat{f}_t \in \mathbb{C}^{c \times h \times w}$, $\hat{f}_{t-1} \in \mathbb{C}^{c \times h \times w}$, and $\hat{f}_{t-2} \in \mathbb{C}^{c \times h \times w}$), which capture the frequency information of the input features. Note that the current spectral features ($\hat{f}_t, \hat{f}_{t-1}$, and $\hat{f}_{t-2}$) are complex numbers and incompatible with the neural

layers. Therefore we concatenate the real and imaginary parts of these complex numbers along the channel dimension respectively and thus obtain three new tensors ($x_t \in \mathbb{R}^{2c \times h \times w}$, $x_{t-1} \in \mathbb{R}^{2c \times h \times w}$, and $x_{t-2} \in \mathbb{R}^{2c \times h \times w}$) with double channels. Afterward, we take the current frame spectral-domain features x_t as the core and fuse the spatial-temporal information from the two auxiliary spectral-domain features (x_{t-1} and x_{t-2}), respectively. Specifically, we first group three features into two groups ($\{x_t, x_{t-1}\}$ and $\{x_t, x_{t-2}\}$) and develop a channel attention function $CA(\cdot)$ to obtain two attention maps. The $CA(\cdot)$ passes an input feature map to a feature normalization, two 1×1 convolution layers $Conv(\cdot)$, a ReLU activation function $\delta(\cdot)$, and a sigmoid function $\sigma(\cdot)$ to compute an attention map. Then, we element-wise multiply the obtained attention map from each group with the input features, and the multiplication results (see y_1 and y_2) are then transformed into complex numbers by splitting them into real and imaginary parts along the channel dimension. After that, inverse FFT (iFFT) operation is employed to transfer the spectral features back to the spatial domain, and then two obtained features at the spatial domain are denoted as z_1 and z_2. Finally, we further element-wisely add z_1 and z_2 and then pass it into a *"BConv"* layer to obtain the output feature o_t of our FFA module. Mathematically, o_t is computed by $o_t = BConv(z_1 + z_2)$, where *"BConv"* contains a 3×3 convolution layer, a group normalization, and a *ReLU* activation function.

3.2   Two-Branch Decoder

After obtaining the frequency features, we introduce a two-branch decoder consisting of a segmentation branch and a localization branch to incorporate temporal features from nearby frames into the current frame. Each branch is built based on the UNet decoder [14] with four convolutional layers. Let d_s^1 and d_s^2 denote the features at the last two layers of the segmentation decoder branch, and d_l^1 and d_l^2 denote the features at the last two layers of the localization decoder branch. Then, we pass d_l^1 at the localization decoder branch to predict a breast lesion localization map. Then, we element-wisely add d_l^1 and d_s^1, and element-wisely add d_l^2 and d_s^2, and pass the addition result into a *"BConv"* convolution layer to predict the segmentation map S_t of the input video frame I_t.

Location Ground Truth. Instead of formulating it as a regression problem, we adopt a likelihood heatmap-based approach to encode the location of breast lesions, since it is more robust to occlusion and motion blur. To do so, we compute a bounding box of the annotated breast lesion segmentation result, and then take the center coordinates of the bounding box. After that, we apply a Gaussian kernel with a standard deviation of 5 on the center coordinates to generate a heatmap, which is taken as the ground truth of the breast lesion localization.

3.3   Location-Based Contrastive Loss

Note that the breast lesion locations of neighboring ultrasound video frames are close, while the breast lesion location distance is large for different ultrasound

Table 2. Quantitative comparisons between our FLA-Net and the state-of-the-art methods on our test set in terms of breast lesion segmentation in ultrasound videos.

Method	image/video	Dice ↑	Jaccard ↑	F1-score ↑	MAE ↓
UNet [14]	image	0.745	0.636	0.777	0.043
UNet++ [19]	image	0.749	0.633	0.780	0.039
TransUNet [4]	image	0.733	0.637	0.784	0.042
SETR [18]	image	0.709	0.588	0.748	0.045
STM [13]	video	0.741	0.634	0.782	0.041
AFB-URR [11]	video	0.750	0.635	0.781	0.038
PNS+ [9]	video	0.754	0.648	0.783	0.036
DPSTT [10]	video	0.755	0.649	0.785	0.036
DCFNet [16]	video	0.762	0.659	0.799	0.037
Our FLA-Net	video	**0.789**	**0.687**	**0.815**	**0.033**

videos, which are often obtained from different patients. Motivated by this, we further devise a location-based contrastive loss to make the breast lesion locations at the same video to be close, while pushing the lesion locations of frames from different videos away. By doing so, we can enhance the breast lesion location prediction in the localization branch. Hence, we devise a **location-based contrastive loss** based on a triplet loss [15], and the definition is given by:

$$\mathcal{L}_{contrastive} = max(MSE(H_t, H_{t-1}) - MSE(H_t, N_t) + \alpha, 0), \quad (1)$$

where α is a margin that is enforced between positive and negative pairs. H_t and H_{t-1} are predicted heatmaps of neighboring frames from the same video. N_t denotes the heatmap of the breast lesion from a frame from another ultrasound video. Hence, the total loss $\mathcal{L}_{total}$ of our network is computed by:

$$\mathcal{L}_{total} = \mathcal{L}_{contrastive} + \lambda_1\mathcal{L}_{MSE}(H_t, G_t^H) + \lambda_2\mathcal{L}_{BCE}(S_t, G_t^S) + \lambda_3\mathcal{L}_{IoU}(S_t, G_t^S), \quad (2)$$

where G_t^H and G_t^S denote the ground truth of the breast lesion segmentation and the breast lesion localization. We empirically set weights $\lambda_1 = \lambda_2 = \lambda_3 = 1$.

4   Experiments and Results

Implementation Details. To initialize the backbone of our network, we pre-trained Res2Net-50 [6] on the ImageNet dataset, while the remaining components of our network were trained from scratch. Prior to inputting the training video frames into the network, we resize them to 352×352 dimensions. Our network is implemented in PyTorch and employs the Adam optimizer with a learning rate of 5×10^{-5}, trained over 100 epochs, and a batch size of 24. Training is conducted on four GeForce RTX 2080 Ti GPUs. For quantitative comparisons, we utilize various metrics, including the Dice similarity coefficient (Dice), Jaccard similarity coefficient (Jaccard), F1-score, and mean absolute error (MAE).
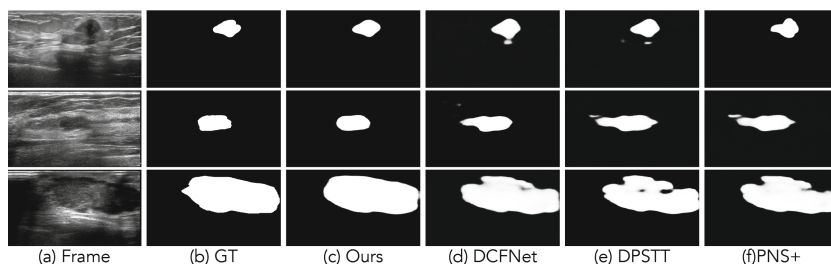
(a) Frame   (b) GT   (c) Ours   (d) DCFNet   (e) DPSTT   (f)PNS+

Fig. 3. Visual comparisons of breast lesion segmentation results produced by our network and state-of-the-art methods. "GT" denotes the ground truth. For more visualization results, please refer to the *supplementary material*.

Table 3. Quantitative comparison results of ablation study experiments.

	FLA	Loc-Branch	Contrastive loss	Dice ↑	Jaccard ↑	F1-score ↑	MAE ↓
Basic	×	×	×	0.747	0.641	0.777	0.037
Basic+FLA	✓	×	×	0.777	0.669	0.806	0.035
Basic+LB	×	✓	×	0.751	0.646	0.781	0.037
Basic+FLA+LB	✓	✓	×	0.780	0.675	0.809	0.034
Our method	✓	✓	✓	**0.789**	**0.687**	**0.815**	**0.033**

4.1 Comparisons with State-of-the-Arts

We conduct a comparative analysis between our network and nine state-of-the-art methods, comprising four image-based methods and five video-based methods. Four image-based methods are UNet [14], UNet++ [19], TransUNet [4], and SETR [18], while five video-based methods are STM [13], AFB-URR [11], PNS+ [9], DPSTT [10], and DCFNet [16]. To ensure a fair and equitable comparison, we acquire the segmentation results of all nine compared methods by utilizing either their publicly available implementations or by implementing them ourselves. Additionally, we retrain these networks on our dataset and fine-tune their network parameters to attain their optimal segmentation performance, enabling accurate and meaningful comparisons.

Quantitative Comparisons. The quantitative results of our network and the nine compared breast lesion segmentation methods are summarized in Table 2. Analysis of the results reveals that, in terms of quantitative metrics, video-based methods generally outperform image-based methods. Among nine compared methods, DCFNet [16] achieves the largest Dice, Jaccard, and F1-score results, while PNS+ [9] and DPSTT [10] have the smallest MAE score. More importantly, our FLA-Net further outperforms DCFNet [16] in terms of Dice, Jaccard, and F1-score metrics, and has a superior MAE performance over PNS+ [9] and DPSTT [10]. Specifically, our FLA-Net improves the Dice score from 0.762 to 0.789, the Jaccard score from 0.659 to 0.687, the F1-score result from 0.799 to 0.815, and the MAE score from 0.036 to 0.033.

Table 4. Quantitative comparison results on different video polyp segmentation datasets. For more quantitative results please refer to the supplementary material.

	Metrics	UNet [14]	UNet++ [19]	ResUNet [7]	ACSNet [17]	PraNet [5]	PNSNet [8]	Ours
CVC-300-TV	Dice ↑	0.639	0.649	0.535	0.738	0.739	0.840	**0.874**
	IoU ↑	0.525	0.539	0.412	0.632	0.645	0.745	**0.789**
	S_α ↑	0.793	0.796	0.703	0.837	0.833	**0.909**	0.907
	E_ϕ ↑	0.826	0.831	0.718	0.871	0.852	0.921	**0.969**
	MAE ↓	0.027	0.024	0.052	0.016	0.016	0.013	**0.010**
CVC-612-V	Dice ↑	0.725	0.684	0.752	0.804	0.869	0.873	**0.885**
	IoU ↑	0.610	0.570	0.648	0.712	0.799	0.800	**0.814**
	S_α ↑	0.826	0.805	0.829	0.847	0.915	**0.923**	0.920
	E_ϕ ↑	0.855	0.830	0.877	0.887	0.936	0.944	**0.963**
	MAE ↓	0.023	0.025	0.023	0.054	0.013	0.012	**0.012**

Qualitative Comparisons. Figure 3 visually presents a comparison of breast lesion segmentation results obtained from our network and three other methods across various input video frames. Apparently, our method accurately segments breast lesions of the input ultrasound video frames, although these target breast lesions have varied sizes and diverse shapes in the input video frames.

4.2   Ablation Study

To evaluate the effectiveness of the major components in our network, we constructed three baseline networks. The first one (denoted as "Basic") removed the localization encoder branch and replaced our FLA modules with a simple feature concatenation and a 1×1 convolutional layer. The second and third baseline networks (named "Basic+FLA" and "Basic+LB") incorporate the FLA module and the localization branch into the basic network, respectively. Table 3 reports the quantitative results of our method and three baseline networks. The superior metric performance of "Basic+FLA" and "Basic+LB" compared to "Basic" clearly indicates that our FLA module and the localization encoder branch effectively enhance the breast lesion segmentation performance in ultrasound videos. Then, the superior performance of "Basic+FLA+LB" over "Basic+FLA" and "Basic+LB" demonstrate that combining our FLA module and the localization encoder branch can incur a more accurate segmentation result. Moreover, our method has larger Dice, Jaccard, F1-score results and a smaller MAE result than "Basic+FLA+LB", which shows that our location-based contrastive loss has its contribution to the success of our video breast lesion segmentation method.

4.3   Generalizability of Our Network

To further evaluate the effectiveness of our FLA-Net, we extend its application to the task of video polyp segmentation. Following the experimental protocol

employed in a recent study on video polyp segmentation [8], we retrain our network and present quantitative results on two benchmark datasets, namely CVC-300-TV [2] and CVC-612-V [3]. Table 4 showcases the Dice, IoU, S_α, E_ϕ, and MAE results achieved by our network in comparison to state-of-the-art methods on these two datasets. Our method demonstrates clear superiority over state-of-the-art methods in terms of Dice, IoU, E_ϕ, and MAE on both the CVC-300-TV and CVC-612-V datasets. Specifically, our method enhances the Dice score from 0.840 to 0.874, the IoU score from 0.745 to 0.789, the E_ϕ score from 0.921 to 0.969, and reduces the MAE score from 0.013 to 0.010 for the CVC-300-TV dataset. Similarly, for the CVC-612-V dataset, our method achieves improvements of 0.012, 0.014, 0.019, and 0 in Dice, IoU, E_ϕ, and MAE scores, respectively. Although our S_α results (0.907 on CVC-300-TV and 0.920 on CVC-612-V) take the 2nd rank, they are very close to the best S_α results, which are 0.909 on CVC-300-TV and 0.923 on CVC-612-V. Hence, the superior metric results obtained by our network clearly demonstrate its ability to accurately segment polyp regions more effectively than state-of-the-art video polyp segmentation methods.

5   Conclusion

In this study, we introduce a novel approach for segmenting breast lesions in ultrasound videos, leveraging a larger dataset consisting of 572 videos containing a total of 34,300 annotated frames. We introduce a frequency and location feature aggregation network that incorporates frequency-based temporal feature learning, an auxiliary prediction of breast lesion location, and a location-based contrastive loss. Our proposed method surpasses existing state-of-the-art techniques in terms of performance on our annotated dataset as well as two publicly available video polyp segmentation datasets. These outcomes serve as compelling evidence for the effectiveness of our approach in achieving accurate breast lesion segmentation in ultrasound videos.

References

1. Bergland, G.D.: A guided tour of the fast Fourier transform. IEEE Spectr. **6**(7), 41–52 (1969)
2. Bernal, J., Sánchez, J., Vilariño, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recogn. **45**(9), 3166–3182 (2012). Best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2011)

3. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput. Med. Imaging Graph. **43**, 99–111 (2015)

4. Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

5. Fan, D.-P., et al.: PraNet: parallel reverse attention network for polyp segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 263–273. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_26

6. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2Net: a new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell. **43**(2), 652–662 (2021)

7. Jha, D., et al.: ResUNet++: an advanced architecture for medical image segmentation. In: 2019 IEEE International Symposium on Multimedia (ISM), pp. 225–2255 (2019)

8. Ji, G.-P., et al.: Progressively normalized self-attention network for video polyp segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 142–152. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_14

9. Ji, G.P., et al.: Video polyp segmentation: a deep learning perspective. Mach. Intell. Res. **19**, 531–549 (2022)

10. Li, J., et al.: Rethinking breast lesion segmentation in ultrasound: a new video dataset and a baseline network. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2022, pp. 391–400. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16440-8_38

11. Liang, Y., Li, X., Jafari, N., Chen, J.: Video object segmentation with adaptive feature bank and uncertain-region refinement. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 3430–3441. Curran Associates, Inc. (2020)

12. Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J., Wang, L.: A new dataset and a baseline model for breast lesion detection in ultrasound videos. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2022, pp. 614–623. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_59

13. 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 (ICCV), October 2019

14. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

15. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015

16. Zhang, M., et al.: 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)

17. Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., Yu, Y.: Adaptive context selection for polyp segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 253–262. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_25

18. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6881–6890, June 2021

19. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: UNet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging **39**(6), 1856–1867 (2020)