Rethinking Breast Lesion Segmentation in Ultrasound: A New Video Dataset and A Baseline Network

Jialu Li[1,2], Qingqing Zheng[2], Mingshuang Li[2], Ping Liu[2], Qiong Wang[2(✉)], Litao Sun[3(✉)], and Lei Zhu[4,5]

[1] University of Chinese Academy of Sciences, Beijing, China
[2] Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
wangqiong@siat.ac.cn
[3] Zhejiang Provincial People's Hospital, Hangzhou, China
litaosun1971@sina.com
[4] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[5] The Hong Kong University of Science and Technology, Hong Kong, China

Abstract. Automatic breast lesion segmentation in ultrasound (US) videos is an essential prerequisite for early diagnosis and treatment. This challenging task remains under-explored due to the lack of availability of annotated US video dataset. Though recent works have achieved better performance in natural video object segmentation by introducing promising Transformer architectures, they still suffer from spatial inconsistency as well as huge computational costs. Therefore, in this paper, we first present a new benchmark dataset designed for US video segmentation. Then, we propose a dynamic parallel spatial-temporal Transformer (DPSTT) to improve the performance of lesion segmentation in US videos with higher computational efficiency. Specifically, the proposed DPSTT disentangles the non-local Transformer along the temporal and spatial dimensions, respectively. The temporal Transformer attends temporal lesion movement on different frames at the same regions, and the spatial Transformer focuses on similar context information between the previous and the current frames. Furthermore, we propose a dynamic selection scheme to effectively sample the most relevant frames from all the past frames, and thus prevent out of memory during inference. Finally, we conduct extensive experiments to evaluate the efficacy of the proposed DPSTT on the new US video benchmark dataset.

1 Introduction

Automatic segmentation of breast lesions in ultrasound (US) video is essential for computer-aided clinical examination and treatment [5]. Compared with the

J. Li and Q. Zheng—Contributed equally to this work.

L. Wang et al. (Eds.): MICCAI 2022, LNCS 13434, pp. 391–400, 2022.
https://doi.org/10.1007/978-3-031-16440-8_38

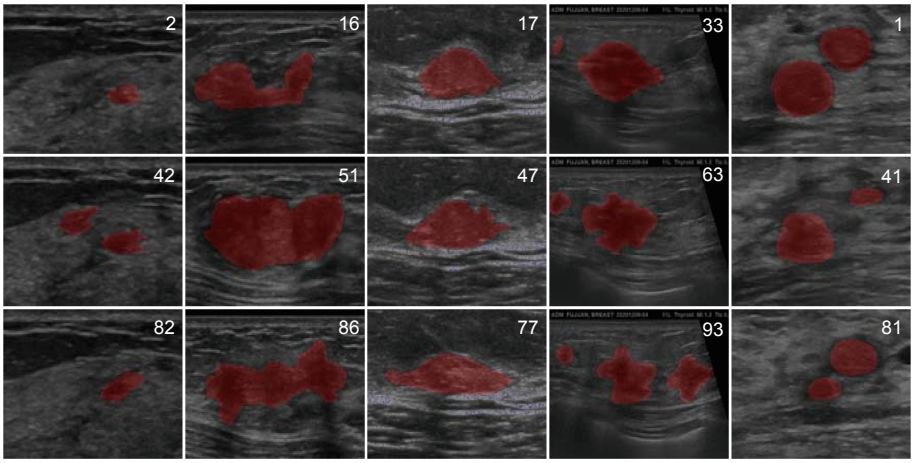

Fig. 1. Each column contains sample frames from a video in the breast lesion segmentation dataset, with high-quality pixel-level annotations marked in red. (Color figure online)

image segmentation, the segmentation in US video is more in line with the practice as it provides additional temporal information for the target object. It can be formulated as a binary labeling problem aiming to automatically segment target lesions in pixel level from a breast US video. This challenging task is rarely explored due to the lack of availability of published annotated US video datasets.

Although many convolutional neural networks (CNN), such as U-Net [9] and its variants [10,13], have achieved outstanding performance on various benchmarks by learning robust representative features for US image segmentation. Directly applying these image segmentation methods to independently process each US video frame may fail to capture temporal context information and result in temporal inconsistency. Recently, Transformers become increasingly popular in video object segmentation tasks [3]. To model long-range relationships, Transformers employ a self-attention mechanism to calculate pairwise similarities between all input units. As a representative, space-time memory (STM) [7] leverages a memory network to read relevant information from a temporal buffer of all preceding frames. The STM performs dense matching in the feature space to capture context information with an unlimited receptive field. However, the non-local property of STM may result in mismatching since that lesions in US videos usually appear in local neighborhoods across memory frames. In addition, the memory would increase linearly with the length of videos during inference, which inevitably brings huge computational costs and may encounter memory overflow.

To tackle the above challenges, we first introduce a new US video dataset with accurate frame-wise annotation in pixel level for breast lesion segmentation; see Fig. 1 for examples. Then, we propose a Dynamic Parallel Spatial-Temporal

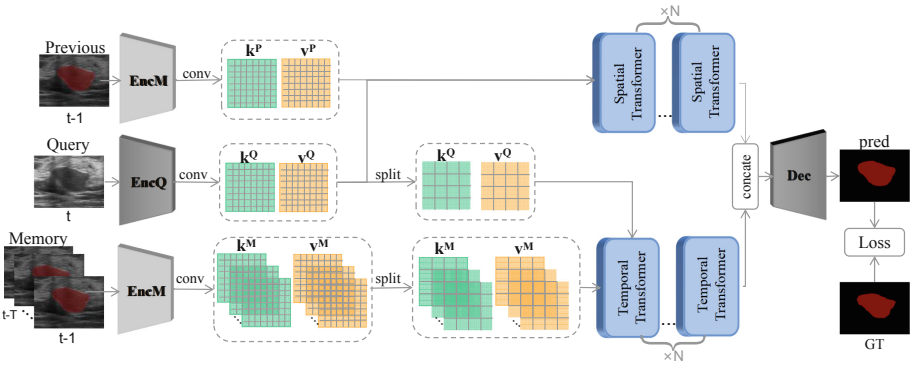

Fig. 2. Overview of the proposed DPSTT framework. Our network consists of two encoders (a memory encoder for the past frames, and a query encoder for the current frame), a parallel pair of spatially- and temporally-decoupled Transformer and a decoder. The memory encoder takes an RGB image and its corresponding lesion mask as input, whilst the query encoder only takes an RGB image. Here we repeat the memory encoder for the previous frame for better visualization.

Transformer (DPSTT) framework for US video segmentation. Specifically, following STM, we first extract pairs of key and value embedding from the current frame and all frames in the memory with a convolution-based encoder. Subsequently, we split the memory module into two parallel temporally- and spatially-decoupled Transformer blocks. In the temporally-decoupled block, the obtained key maps are spatially divided into multiple non-overlapped patches, and the attention is only calculated in the same regions between embedding of the current frame and those of memory frames. Such a temporal operation makes the modeling of pixel movements of breast lesions easier. By contrast, the spatially-decoupled block calculates the attention between the embedding of the current frame and that of the previous one in a non-local manner, which models the global similarity of stationary background texture between two adjacent frames. Moreover, to prevent unlimited growth of memory during inference, we also develop a non-uniform adaptive memory selection scheme to dynamically update the frames in the memory based on the similarity metric. In summary, the contributions of our method are threefold: (1) We are the first to present an annotated benchmark dataset specifically designed for the task of breast lesion segmentation in US videos, which would promote the progress of the medical video process. (2) We propose a Dynamic Parallel Spatial Temporal Transformer (DPSTT) framework for US video segmentation to improve lesion segmentation performance with higher computational efficiency. (3) We have conducted extensive experiments to evaluate the proposed DPSTT. Experimental results demonstrate that our method outperforms state of the arts by a large margin.

2 Method

The overall framework of the proposed DPSTT is shown in Fig. 2. Given a US video sequence, we regard the current frame as the *query* frame, the past frames with annotated object masks as the *memory* frames. During the video segmentation process, both memory frames and the query frame are first encoded into pairs of key and value maps through the memory encoder and the query encoder, respectively. Different from STM that constructs a global memory read module over the video space, we disentangle the non-local attention into two parallel lightweight modules along the spatial and temporal dimensions. The keys and values further go through the spatially-decoupled and temporally-decoupled Transformers. Specifically, the spatial Transformer takes the keys and values from the query and the previous frames to extract the global background context information, while the temporal Transformer takes the keys and values at the same local regions from the query and memory frames to aggregate the temporal movement of target objects at the same time. The outputs of the spatial and temporal Transformers are finally sent to the decoder, which estimates the target mask for the query frame.

2.1 Query and Memory Encoder

Both the query and memory encoders share the same structure except for the input. Similar to STM, we utilize the ResNet50 [4] as the backbone network and modify the first convolutional layer to take a 4-channel input for the memory encoder. Then two parallel convolutional layers are utilized to further embed the backbone network output into a pair of key and value maps by reducing its channel size to 1/8 and 1/2, respectively. We denote by $\mathbf{k}^Q \in \mathbb{R}^{H\times W\times C/8}$ and $\mathbf{v}^Q \in \mathbb{R}^{H\times W\times C/2}$ the key and the value maps for the query frame, where H is the height, W is the width and C represents the channel size of the feature map. Similarly, each individual of T memory frames ($T \geq 1$) is independently embedded into key and value maps. The resulting key and value maps are represented as $\mathbf{k}^M \in \mathbb{R}^{T\times H\times W\times C/8}$ and $\mathbf{v}^M \in \mathbb{R}^{T\times H\times W\times C/2}$. For ease of description, we also denote the corresponding key and value maps of the previous frame by $\mathbf{k}^P$ and $\mathbf{v}^P$, which have the same resolution as $\mathbf{k}^Q$ and $\mathbf{v}^Q$.

2.2 Parallel Spatial Temporal Transformer

Different from the memory read module in STM that simultaneously processes similarity matching between all pixels of the query frame and memory frames, we disentangle this expensive module into two much easier components: a temporally-decoupled Transformer for extracting local features along the temporal dimension, and a spatially-decoupled Transformer block for capturing global features between the query frame and its previous frame in a non-local manner.

For the temporal Transformer, given the memory key $\mathbf{k}^M$ and the query key $\mathbf{k}^Q$, we split them into s^2 non-overlapped patches along both height and width dimensions. Each region is represented by $\mathbf{k}_{ij}^{M_k} \in \mathbb{R}^{H/s\times W/s\times C/8}, k \in [1, T]$ for

the memory and $\mathbf{k}_{ij}^{Q} \in \mathbb{R}^{H/s \times W/s \times C/8}$ for the query respectively, where $i, j \in [1, s]$ denote the index of the local region. We then group the local memory regions by temporal dimension, *i.e.*, $P_{ij}^{M} = \{\mathbf{k}_{ij}^{M_1}, \cdots, \mathbf{k}_{ij}^{M_T}\}$. Then the temporal Transformer measures the local similarity with:

$$f(P_{ij}^{M_k}, \mathbf{k}_{ij}^{Q}) = \text{Softmax}\big(\exp(P_{ij}^{M_k} \otimes \mathbf{k}_{ij}^{Q})\big), \tag{1}$$

where $\otimes$ denotes the dot product. With the soft weights, the memory values are subsequently retrieved by a weighted summation as follows:

$$\mathbf{v}_{ij}^{T} = \sum_{k=1}^{T} f(P_{ij}^{M_k}, \mathbf{k}_{ij}^{Q}) \mathbf{v}_{ij}^{M_k}. \tag{2}$$

The resulting $\mathbf{v}_{ij}^{T}$ concatenated with the query value at the same location, is further organized into a new tensor according to its location index to produce the temporal Transformer output $\mathbf{y}^{T}$. By doing so, continuous movements of the target object in a smaller spatial region can be detected without the disturbance of redundant temporal features.

For the spatial Transformer, we assume the previous frame has less movement or appearance difference compared with the query frame. The previous frame with its estimated mask would help provide coarse guidance for the query frame. Therefore, the similarity matching between the previous and the query frame is performed in a non-local manner with

$$f(\mathbf{k}^{Q}, \mathbf{k}^{P}) = \text{Softmax}\big(\exp(\mathbf{k}^{Q} \otimes \mathbf{k}^{P})\big). \tag{3}$$

Then the output of the spatial Transformer is generated by

$$\mathbf{y}^{S} = [\mathbf{v}^{Q}, f(\mathbf{k}^{Q}, \mathbf{k}^{P}) \mathbf{v}^{P}]. \tag{4}$$

In this way, the spatial Transformer pays more attention to the global static background context information. Finally, these two decoupled spatial and temporal Transformers are calculated in parallel and their outputs $\mathbf{y}^{S}$ and $\mathbf{y}^{T}$ are concatenated and further refined by a convolutional operation.

2.3 Decoder

The decoder takes the refined output of the decoupled spatial and temporal Transformers to estimate the lesion mask for the query frame. We follow the refinement module in [7] to build the decoder, which upscales the feature map gradually by a set of residual convolutional blocks. Finally, We minimize the binary cross-entropy(BCE) loss and the dice loss between the object masks $\hat{Y}$ and the ground truth labels Y.

2.4 Dynamic Memory Selection

Though spatial and temporal Transformers benefit from storing enough information in the memory frames, storing all the past frames is impossible and may

Table 1. Quantitative comparison with different methods on the proposed dataset.

Methods	Jaccard	Dice	Precision	Recall	FPS
UNet [9]	62.47 ± 0.53	73.03 ± 0.36	79.46 ± 0.20	72.72 ± 0.45	**88.18**
UNet++ [13]	61.24 ± 0.73	71.79 ± 0.53	82.80 ± 0.04	68.84 ± 1.09	40.9
TransUNet [2]	53.58 ± 0.37	65.47 ± 0.21	71.67 ± 0.13	66.82 ± 0.20	65.1
SETR [12]	54.80 ± 0.68	66.49 ± 0.59	75.33 ± 0.15	66.43 ± 1.04	21.61
OSVOS [8]	56.74 ± 0.59	70.98 ± 0.33	77.78 ± 0.92	64.04 ± 0.98	27.25
ViViT [1]	54.46 ± 0.32	67.39 ± 0.29	75.54 ± 0.03	66.83 ± 0.59	24.33
STM [7]	68.58 ± 0.56	78.62 ± 0.43	82.01 ± 0.35	79.10 ± 0.44	23.17
AFB-URR [6]	70.34 ± 0.25	80.18 ± 0.15	80.08 ± 0.32	$\mathbf{85.91 \pm 0.15}$	11.84
Ours	$\mathbf{73.64 \pm 0.18}$	$\mathbf{82.55 \pm 0.20}$	$\mathbf{83.89 \pm 0.13}$	84.55 ± 0.29	30.5

lead to memory overflow. To eliminate unnecessary features, we propose a simple yet effective selection mechanism to dynamically update the memory frames. We maintain a fixed K memory frames for segmenting the t^{th} query frame if $t > K$, or all of the past frames as memory frames if $t \leq K$. Then we update the memory buffer by selecting the most K relevant frames. For example, assume that we have memory frames M for segmenting the t^{th} frame ($t > K$), when moving forward to the $(t+1)^{th}$ frame, we adopt the cosine metric and sort the resulting similarity values with

$$\text{Sort}\{Cos(\mathbf{k}^{Q_{t+1}}, \mathbf{k}^{M_k}), Cos(\mathbf{k}^{Q_{t+1}}, \mathbf{k}^{M_t})\}, k \in [1, K]. \tag{5}$$

Then the memory frames can be updated by adding the t^{th} frame at the tail and removing the one with the least similarity value. It is noteworthy that our dynamic memory selection speeds up the temporal attention calculation, and performs online adaptation without additional training.

Time Complexity. With such a pipeline, we significantly reduce the computational complexity of memory read module in STM from $\mathcal{O}(TH^2W^2C)$ into $\mathcal{O}(KH^2W^2C/(s^4))$ for the temporally-decoupled block and $\mathcal{O}(H^2W^2C)$ for the spatially-decoupled block. Although $T = 3$ is chosen in the training process, T would increase linearly with the video length during inference, which would be much larger than a predetermined K. In addition, the computation of the temporal Transformer would be more efficient when s becomes larger.

3 Experiments

3.1 Dataset and Implementation

Here we describe the newly collected dataset, specifically designed for the task of breast lesion segmentation in US videos. Sample frames of the breast US videos

are shown in Fig. 1. The breast US dataset comprises 63 video sequences, one video sequence per person, 4619 frames annotated with pixel-level ground truth by experts. These videos are collected from different US devices and their spatial resolution varies from 580×600 to 600×800. To ease training, we further crop the video sequences to a spatial resolution of 300×200. For quantitative comparison, we employ several widely used segmentation evaluation metrics, namely, Jaccard similarity coefficient (Jaccard), Dice similarity coefficient (Dice), Precision and Recall; see [11] for their definitions. Moreover, we adopt five-fold cross-validation on our dataset to statistically test different video segmentation methods.

We implement our network using the PyTorch framework with an NVIDIA RTX 3090 graphics card. In our experiments, all the input US frames are empirically resized to 240×240 and the training epoch is set to 100. During training, we sample $T (T = 3)$ temporally ordered frames with random skip N frames $(N \leq 5)$ from a US video. We set the batch size to 4 and learning rate to $1e-4$. We use the binary cross-entropy(BCE) loss and the dice loss with the weight of 0.5 and 0.5 during the training process. For the temporal Transformer, we set s to be 2. During inference, when the size of memory frames exceeds K $(K = 10)$, the dynamic selection mechanism is activated to eliminate the redundant frame for the memory.

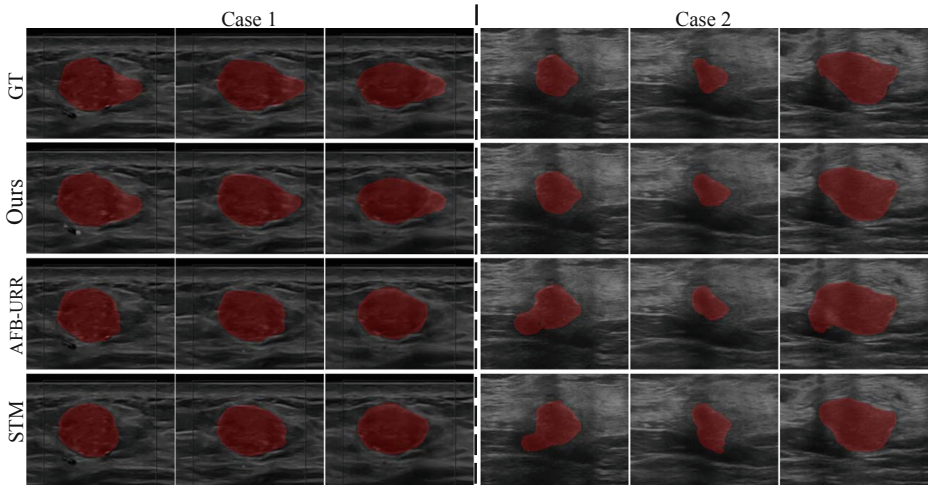

Fig. 3. Visual comparison with competitive video-based methods on two breast lesion cases.

3.2 Comparison with State-of-the-Art Methods

Quantitative Comparisons. As shown in Table 1, we qualitatively compare our method with state-of-the-art methods, including image-based segmentation methods (UNet [9], UNet++ [13], TransUNet [2] and SERT [12]) and video-based

Table 2. Ablation study on different transformer combination strategies. T denotes the temporal block and S is the spatial block. $\times N$ denotes repeating N times.

Stacking strategies	Jaccard	Dice	Precision	Recall
S(x1)	70.92 ± 0.15	80.07 ± 0.19	81.63 ± 0.25	82.77 ± 0.66
T(x1)	71.09 ± 0.15	80.24 ± 0.21	83.61 ± 0.26	80.88 ± 0.43
S-T(x1)	72.64 ± 0.18	81.58 ± 0.23	83.63 ± 0.11	82.99 ± 0.47
T-S(x1)	72.86 ± 0.18	81.86 ± 0.21	82.75 ± 0.08	84.02 ± 0.44
T||S(x1)	$\mathbf{73.64 \pm 0.18}$	$\mathbf{82.55 \pm 0.20}$	$\mathbf{83.89 \pm 0.13}$	$\mathbf{84.55 \pm 0.29}$
S-T(x3)	72.03 ± 0.18	81.22 ± 0.19	82.69 ± 0.20	83.07 ± 0.25
T-S(x3)	72.72 ± 0.22	81.68 ± 0.29	83.18 ± 0.09	83.67 ± 0.61
T||S(x3)	72.15 ± 0.29	81.64 ± 0.24	83.11 ± 0.47	83.10 ± 0.15

segmentation methods (OSVOS [8], ViViT [1], STM [7], AFB-URR [6]). From the results, we can observe that the video-based methods are prone to outperform image-based methods with higher evaluation scores, which demonstrates that leveraging temporal information provides promising benefits for breast lesion segmentation in US videos. More importantly, among all video-based segmentation methods, our DPSTT has achieved the highest Jaccard score of 73.64 and the Dice score of 82.55. It indicates that our method, combined with a CNN-based encoder and spatial-temporal Transformers, is able to simultaneously learn both high- and low-level cues and thus achieves significant improvements over those pure Transformer methods, such as SERT and ViViT. Table 1 also reports the inference speed performance of different methods. Due to the parallel operation of the decoupled Transformers, our method reduces much redundant computation and thus runs the fastest compared with other video-based approaches.

Qualitative Comparisons. Figure 3 visualizes the qualitative comparison of lesion masks among different video segmentation methods. We can observe that our method can provide more precise masks than STM and AFB-URR with more consistent boundaries. This is because our dynamic selection mechanism provides the most relevant memory and preserves the spatial consistency.

3.3 Ablation Study

The Effect of Transformers. We evaluate the effect of different Transformers by removing the spatial and temporal blocks separately in Table 2. It shows that the combination of both modules consistently results in better performance. This is because any decoupled Transformer can't simultaneously capture both stationary texture and moving information. We further compare different stacking strategies in the row 3–5. It shows that stacking such two different Transformers in parallel performs better than in an interweaving way, no matter starting from spatial or temporal blocks. Moreover, it is also observed that only one parallel temporal and spatial blocks are good enough to capture representative features.

Table 3. Ablation study on different memory selection strategies.

Sample strategies	Jaccard	Dice	Precision	Recall
Skip memory	73.06 ± 0.19	82.10 ± 0.21	83.61 ± 0.16	84.13 ± 0.30
Random memory	72.76 ± 0.16	81.94 ± 0.19	83.04 ± 0.09	83.95 ± 0.30
Dynamic memory	$\mathbf{73.64 \pm 0.18}$	$\mathbf{82.55 \pm 0.20}$	$\mathbf{83.89 \pm 0.13}$	$\mathbf{84.55 \pm 0.29}$

The Effect of Dynamic Memory. We investigate different memory selection strategies by comparing the segmentation performance with skip memory (every five frames) in STM, random memory of fixed size as well as our dynamic memory in Table 3. The results show that the random memory performs worst than the other two strategies. This phenomenon verifies our assumption that good enough memory can provide benefits for segmentation performance.

4 Conclusion

In this paper, we present the first pixel-wise annotated benchmark dataset for breast lesion segmentation in US videos. Then a Dynamic Parallel Spatial-Temporal Transformer framework is proposed for US video segmentation. Moreover, an efficient dynamic memory selection is further developed based on the similarity metric to prevent memory overflow. Finally, we conduct extensive experiments to evaluate the efficacy of our method.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 12026604, No. 62072452 and No. 61902275), the Regional Joint Fund of Guangdong under Grant (No. 2021B1515120011), the Key Fundamental Research Program of Shenzhen under Grant (No. JCYJ20200109115627045 and No. JCYJ20200109114233670) and in part by Pazhou Lab, Guangzhou 510320, China.

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: a video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6836–6846 (2021)
2. Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation. CoRR, abs/2102.04306 (2021)
3. Duke, B., Ahmed, A., Wolf, C., Aarabi, P., Taylor, G.W.: SSTVOS: sparse spatiotemporal transformers for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5912–5921 (2021)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, June 2016
5. Huang, Q., Huang, Y., Luo, Y., Yuan, F., Li, X.: Segmentation of breast ultrasound image with semantic classification of superpixels. Med. Image Anal. **61**, 101657 (2020)

6. Liang, Y., Li, X., Jafari, N., Chen, Q.: Video object segmentation with adaptive feature bank and uncertain-region refinement. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2020)
7. 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), pp. 9226–9235 (2019)
8. Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., Sorkine-Hornung, A.: Learning video object segmentation from static images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2663–2672 (2017)
9. 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
10. Schlemper, J.: Attention gated networks: learning to leverage salient regions in medical images. Med. Image Anal. **53**, 197–207 (2019)
11. Wang, Y., et al.: Deep attentional features for prostate segmentation in ultrasound. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 523–530. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_60
12. 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 (2021)
13. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1