Ultrasound Video Segmentation with Adaptive Temporal Memory

He Zhao[1(✉)], Qianhui Men[1], Alexander Gleed[1], Aris T. Papageorghiou[2], and J. Alison Noble[1]

[1] Institute of Biomedical Engineering, University of Oxford, Oxford, UK
he.zhao@eng.ox.ac.uk
[2] Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK

Abstract. Automated segmentation of anatomical structures in fetal ultrasound video is challenging due to the highly diverse appearance of anatomies and image quality. In this paper, we propose an ultrasound video anatomy segmentation approach to iteratively memorise and segment incoming video frames, which is suitable for online segmentation. This is achieved by a spatio-temporal model that utilizes an adaptive memory bank to store the segmentation history of preceding frames to assist the current frame segmentation. The memory is updated adaptively using a skip gate mechanism based on segmentation confidence, preserving only high-confidence predictions for future use. We evaluate our approach and related state-of-the-art methods on a clinical dataset. The experimental results demonstrate that our method achieves superior performance with an F1 score of 84.83%. Visually, the use of adaptive temporal memory also aids in reducing error accumulation during video segmentation.

1 Introduction

In obstetric ultrasound, it is crucial to efficiently identify and segment various anatomical structures in the fetomaternal environment including the placenta and maternal bladder. This is because the mutual position between these two anatomies can indicate obstetric complications and thus inform the safest mode of delivery [11,15]. Such anatomy location and morphology analysis typically involve a large amount of manual effort, which is expensive due to the required expertise and is prone to inter- and intra-observer variation. Automated segmentation of the placenta and bladder can provide valuable information for computer-aided diagnosis. However, it is difficult to define the boundaries of such maternal anatomies because of high variations in shape and low contrast of the ultrasound video.

Several works [5,12,14,18,20,21] have attempted automated segmentation of maternal anatomies. A typical approach is to employ a 2D network for single image segmentation [7], such as U-Net [13] and its variations. For instance, four U-Net-based networks are used in [21] to segment a placenta image from

B. Kainz et al. (Eds.): ASMUS 2023, LNCS 14337, pp. 3–12, 2023.
https://doi.org/10.1007/978-3-031-44521-7_1

multiple views of 3D ultrasound volumes. They further propose a multi-task learning approach [20] of placenta position prediction to complement the placenta segmentation task. A coarse-to-fine segmentation pipeline is introduced in [5], where the initial anatomy segmentation is generated by a U-Net model and refined by conditional random field as a recurrent neural network. In [18], a multi-object segmentation network is proposed to segment anatomies in an ultrasound volume. The current ultrasound segmentation methods focus solely on individual timestamp and do not take into account the temporal relationship, leading to inadequate and inconsistent segmentation. Recently, a video-based segmentation method is proposed [3] to recognize breast lesions, where a 3D convolutional network with the additional temporal dimension over images is modelled to reconstruct the segmentation from a pseudo mask. However, their method requires the entire video to be observed, which makes it not applicable to online video streams.

In this paper, we propose a video-based approach for online segmentation by modelling the temporal dynamic behaviour of ultrasound video. We follow the protocol of one-shot video-object segmentation (OS-VOS) [1]: given only the first-frame annotation, the model conducts a closed-loop prediction that automatically segments subsequent frames. Our approach utilizes a memory network [10,19] to store the temporal information of ultrasound video. Inspired by [2] to update states in RNN, we propose a skip gate mechanism on the memory network, and the memory is further selected by a scoring function [8]. Then, a combined pixel and region loss encourages the model to consider both local and regional information, thus facilitating segmentation of accurate shape and boundary. The contributions of our paper are summarized as follows: 1) We propose a spatio-temporal model for ultrasound video segmentation with a memory bank, which provides new insight for video segmentation by effectively utilizing information from preceding frames. 2) An adaptive temporal memory module is proposed to update the memory bank with a skip gate, which reduces the error accumulation and maintains temporal consistency. 3) A combined pixel and region loss is proposed to learn the shape and boundaries of segmented regions. An investigation of our approach on an unseen anatomy, *i.e.,* fetal head, illustrates the generalisability of our approach to other anatomical structures.

2 Method

Our goal is to segment multiple objects in incoming video frames by referring to the first-frame segmentation. The idea of our model is to use the historical sequential information retrieved from an adaptive temporal memory bank to assist in accurate segmentation of the current frame. The current frame $I_t \in \mathbb{R}^{W \times H \times 3}$ and the preceding frames $\boldsymbol{I} = \{I_i | i = 1, ..., t-1\}$ are considered as query and memory, respectively. As shown in Fig. 1, the complete framework consists of three parts: a spatial feature extraction module $\mathcal{F}$ with two encoders E_Q and E_M used for extracting the spatial representations of query and memory frames; an adaptive temporal memory module which updates from temporal

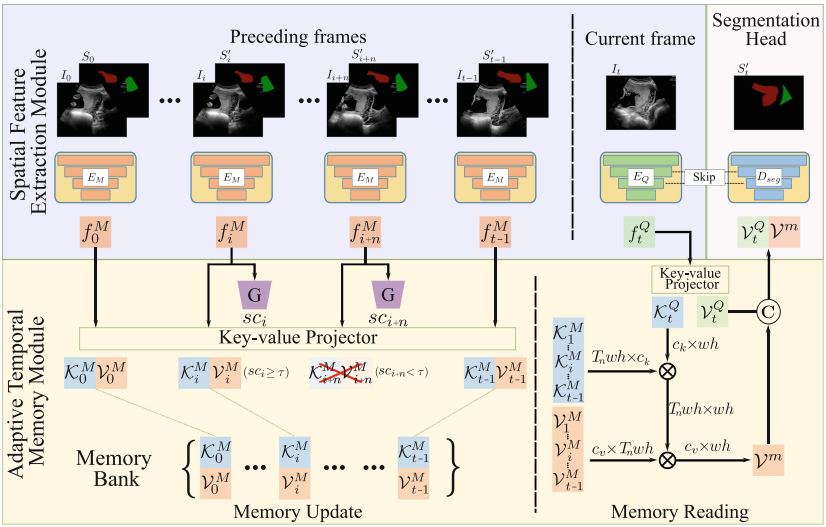

Fig. 1. Flowchart of our architecture. It has three components: the feature extraction module for spatial feature learning by encoder E_M and E_Q; the adaptive temporal memory module controlled by a skip gate mechanism for memory update and memory reading; and a segmentation head with skip-connection decoder D_{seg}. $\otimes$ denotes matrix inner-product and © represents the concatenation operation.

dependencies of memory frames $\boldsymbol{I}$ and their predicted masks $\boldsymbol{S'}$; and a segmentation head, fusing retrieved memory embedding $\mathcal{V}^m$ and query embedding $\mathcal{V}_t^Q$ to predict the current frame segmentation.

2.1 Spatial Feature Extraction Module

The memory encoder E_M takes both the preceding frame I_i and its predicted segmentation mask S_i' as input, and outputs spatial features $f^M = \{E_M(I_i, S_i')|i = 0, \ldots, t-1\}$, where S_i' helps to identify the spatial features of related targets from the background. Different from E_M, the query encoder E_Q only takes the current frame I_t as input and produces a feature map $f_t^Q = E_Q(I_t)$. Other than an extraction layer in E_M to deal with the segmentation input, E_M and E_Q share the same model structure of a ResNet-50 [6] as the feature extraction backbone.

2.2 Adaptive Temporal Memory Module

Memory Construction and Reading. After the spatial feature extraction module $\mathcal{F}$, each of the spatial features f^M from memory and f_t^Q from the query is embedded into a *key* matrix $\mathcal{K} \in \mathbb{R}^{w \times h \times c_k}$ and a *value* matrix $\mathcal{V} \in \mathbb{R}^{w \times h \times c_v}$ by two convolutional layers, where c_k and c_v are the corresponding embedded channel dimensions, respectively. $\mathcal{K}$ is learned to retrieve the relevant feature embedding from the spatial information stored in $\mathcal{V}$. The query value $\mathcal{V}_t^Q$ focuses on

the object appearance information at the current time t. The memory value $\mathcal{V}_i^M$ learns the relationship between frame and object segmentation. Each key-value pair from a preceding frame is stored in the memory bank $\mathcal{M} = \{(\mathcal{K}_i^M, \mathcal{V}_i^M)\}$ with size $|\mathcal{M}| = T_n$, which records the segmentation history and encodes the object motion across the preceding frames that is useful for subsequent frame segmentation. The memory embedding for the current frame is retrieved by the similarity between the query key $\mathcal{K}^Q$ and the memory key $\mathcal{K}^M$, *i.e.,* $\mathcal{V}^m = W\mathcal{V}^M$, where the entry of W is defined as:

$$\omega_j = \frac{\exp(z\mathcal{K}_t^Q \cdot \mathcal{K}_j^M)}{\sum_l \exp(z\mathcal{K}_t^Q \cdot \mathcal{K}_l^M)}. \tag{1}$$

Here z is the scaling factor that is set to $\frac{1}{\sqrt{c_k}}$ [17].

Skip Gate for Memory Update. As the memory bank consists of information from each preceding frame, the segmentation error will be accumulated during this process. To alleviate this problem, we propose to adaptively update the memory bank with a skip gate mechanism. The key insight is to introduce a score function to control the temporal information flow that preserves the memories only with frames given a high segmentation confidence. The skip gate G for the memory update is implemented by a trainable convolutional network to predict a confidence score sc from the memory feature f_i^M at time i with $sc_i = G(f_i^M) = G(E_M(I_i, S_i'))$. Here, G consists of three convolutional layers, two fully-connected layers, and a sigmoid function such that the predicted score is within the range of $[0, 1]$. The role of G is as a regression function to predict the confidence level of the correspondence between the frame and its predicted segmentation. Only the frames with a confidence score larger than a predefined threshold τ are used to update the memory bank. During inference, the proposed skip gate also reduces the burden of the memory bank to enable fast segmentation.

2.3 Segmentation

Segmentation Head. During segmentation, the predicted mask is generated by decoder D_{seg} from both the query value $\mathcal{V}_t^Q$ at the current frame t and the retrieved memory value $\mathcal{V}^m$ of the past frames $i < t$. The decoder is built based on three ResBlock with skip connections. The initial ResBlock merges $\mathcal{V}^m$ and $\mathcal{V}_t^Q$ to extract comprehensive spatial information for the current segmentation, and the decoder output is interpolated to the size of I_t as the predicted segmentation mask.

Training. The overall objective function $\mathcal{L}$ is a combination of two loss components: a segmentation loss $\mathcal{L}_{seg}$ and a confidence loss $\mathcal{L}_{sc}$, *i.e.,* $\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_{sc}$. The detailed construction of each loss component is explained next.

Segmentation Loss. We consider both the local and global errors in the generated segmentation mask with a pixel loss and a region loss, respectively. The pixel loss is constructed from the cross entropy between the corresponding segmented pixels, which aims to identify each pixel independently. The region loss is IoU-based to minimize the mismatched area between two segmentation masks. Additionally, the inclusion of region loss also helps to alleviate the foreground and background imbalance present in ultrasound images. Combining the two losses, the segmentation loss $\mathcal{L}_{seg}$ is given by

$$\mathcal{L}_{seg} = \underbrace{-\sum_{n}(s_n log(s'_n) + (1 - s_n) log(1 - s'_n))}_{\text{pixel loss}} + \underbrace{\frac{\sum_n s_n s'_n}{\sum_n (s_n + s'_n - s_n s'_n)}}_{\text{region loss}}, \quad (2)$$

where s_n and s'_n stands for the nth pixel in the segmentation ground truth S and prediction S', respectively.

Confidence Loss. The skip gate G predicts the segmentation confidence based on the image and segmentation features. To train network G, the ground truth of confidence score of each frame segmentation is defined as the IoU between the segmentation prediction and its corresponding ground truth. The segmentation confidence loss $\mathcal{L}_{sc}$ for optimizing the skip gate is defined as:

$$\mathcal{L}_{sc} = (sc - IoU(S, S'))^2 \quad (3)$$

where $sc = G(E_M(I, S'))$ is the output from skip gate network.

3 Experiments and Results

The dataset used in this paper consisted of 15 ultrasound video scans from the CALOPUS project [16] that used a U-shaped video sweep protocol to scan from the maternal right to left over the top of the pelvis. The videos were randomly split into 11 for training and 4 for test. Each video is approximately 20 s containing around 200 frames after downsampling, with a video frame size of 1008×784. For each video frame, a manual segmentation annotation of the placenta and maternal bladder was available as ground truth. To increase the robustness of automated segmentation, we randomly selected three frames in temporal order with resized shape of 448×448 as one training sample. The model was trained with an Adam optimizer for 200 epochs with a decayed learning rate of $2e^{-5}$. During inference, a whole video sequence is iteratively segmented given the manual annotation of the starting frame as reference.

3.1 Evaluations

We compared our method with three image-based models which only use the current frame to predict the segmentation mask: U-Net [13], ResNet34 [6] and

Table 1. Quantitative performance of image- and video-based segmentation methods evaluated by the F1-score, accuracy, Jaccard Index, Hausdorff distance, and contour accuracy.

Protocol	Method	F1-score ↑	Accuracy ↑	Jaccard ↑	Hausdorff ↓	Contour ↑
Image-based	U-Net [13]	62.51	97.67	52.93	83.60	26.57
	ResNet34 [6]	64.25	96.48	50.69	97.24	21.04
	TransU-Net [4]	68.00	97.52	59.15	77.84	34.24
Video-based	STM [10]	82.50	99.20	73.15	46.16	39.29
	Ours	**84.83**	**99.39**	**76.25**	**33.73**	**40.03**

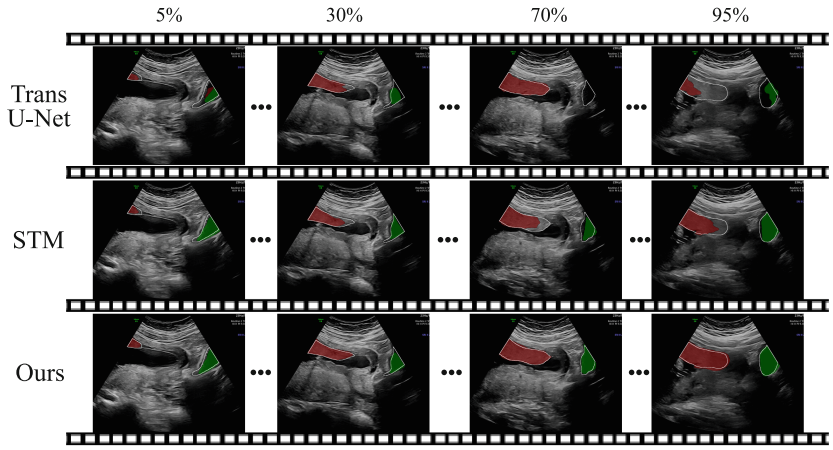

Fig. 2. Qualitative results of video segmentation. Red: placenta; Green: bladder. The white boundary line is the segmentation ground truth of each object. Different frame positions are shown as the percentage of the video length. (Color figure online)

TransU-Net [4]; and a video-based segmentation model under OS-VOS protocol [1] – Spatio-temporal Memory Network (STM) [10]. Table 1 compares these methods with our approach using five segmentation metrics: F1-score, accuracy, Jaccard Index, Hausdorff distance, and contour accuracy. Among those metrics, Hausdorff distance and contour accuracy inform about the object boundary and shape which are important in our clinical application.

First, we observe that the two video-based methods achieve higher overall scores than image-based methods, which suggests that the information from prior video frames is helpful during the video segmentation process. Compared with video-based STM, our approach with adaptive temporal memory achieves superior performance with an improvement of F1-score by 2.8%. Our approach also achieves the lowest Hausdorff distance (33.73 pixels) and the highest contour accuracy score (40.03%). The accurate shapes and border regions of placenta and bladder are important indicators for fetal diagnosis. For instance, the distance between the lower boundary of placenta and the bottom of bladder can be used

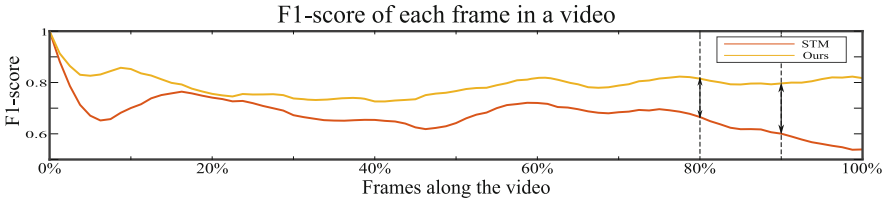

Fig. 3. The quantitative comparison (F1-score) between STM and our approach with adaptive memory module on each frame along the video.

Table 2. Ablation studies for different experimental settings.

Method		F1-score ↑	Accuracy ↑	Jaccard ↑	Hausdorff ↓	Contour ↑
Latest frame only		73.18	99.18	64.52	43.94	32.20
First & latest frames		78.21	99.30	69.70	37.93	34.35
Preceding frames	w/o ATM	82.50	99.20	73.15	46.16	39.29
	w/o region loss	81.45	99.37	72.85	35.53	37.42
	Ours	**84.83**	**99.39**	**76.25**	**33.73**	**40.03**

to differentiate normal and abnormally-located placentae [9]. Figure 2 shows typical visual segmentation results for TransU-Net, STM, and our approach. The TransU-Net segmentation results are less consistent and less accurate compared to the video-based methods, as it only considers the current frame appearance and ignores segmentation history. Within the video-based methods, the segmentation error for STM quickly accumulates as prediction progresses - *c.f.* the result at 70% and 95% of the whole video. Figure 3 illustrates the F1-score of each frame along a video. It can be observed that the model without adaptive memory (*i.e.,* STM) experiences a significant decrease in performance for the later frames in the video. By keeping only the memory with high segmentation confidence in the temporal domain, our approach does not suffer the same performance degradation.

3.2 Ablation Study

Temporal Memory Bank. We first analyze the influence of using temporal information. Four scenarios are compared in this experiment: 1) only the latest frame (the frame before current frame) used as memory; 2) the first annotated frame and the latest frame as memory; 3) all preceding frames as memory without adaptive temporal memory (denoted as *w/o ATM*), and 4) preceding frames with adaptive temporal memory (*Ours*). The quantitative and qualitative results are shown in Table 2 and Fig. 4, respectively. The model with only the latest frame as memory produces the lowest F1-score (73.18%) and fails to segment the bladder. The performance increases to 78.21% by adding the first frame into the memory, since the first annotation is a key reference to inform the model with the position of the segmentation. With more preceding frames

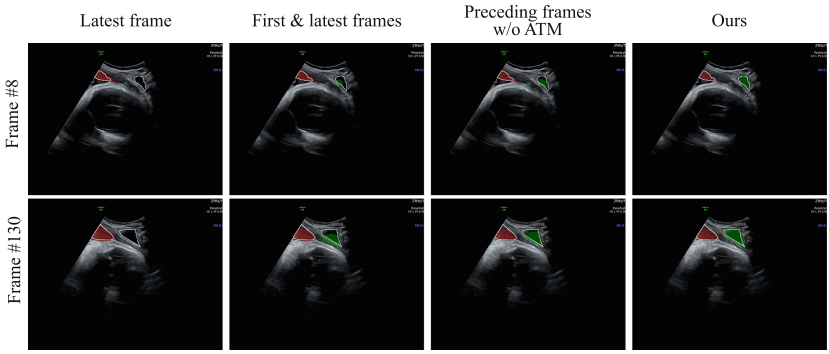

Fig. 4. Qualitative comparisons under different memory settings.

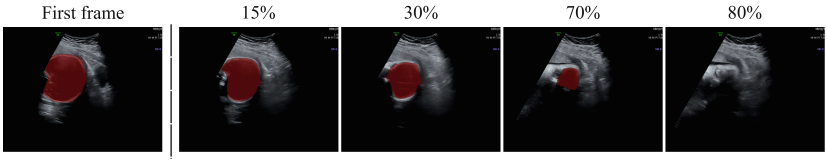

Fig. 5. Visual results on unseen anatomy (fetal head) segmentation.

included (as in *w/o ATM*), the temporal dependencies and the motion of objects in video are modelled in the memory. This allows the model to recall and leverage information from previous frames to generate a plausible prediction on the current frame. Adding the adaptive temporal memory mechanism (as in *Ours*) further boosts segmentation performance, which utilizes the skip gate to encourage incorrectly segmented frames to be discarded. This enables the model to self-check and rectify its own errors, resulting in a more accurate segmentation.

Pixel and Region Loss. We tested the model with different loss terms. The results are reported in the last two rows of Table 2. The model with both pixel and region loss (shown as *Ours*) achieves stronger metric scores in terms of regional evaluation, *i.e.,* Jaccard Index and contour accuracy. This demonstrates that the region loss encourages the model to pay more attention to the whole area and thus anatomical structure, resulting in more precise object boundaries.

Sensitivity of First-Frame Annotation. It is of interest to qualify the model robustness towards the variations of the first-frame annotation. For a test video, first-frame masks of placenta and bladder from three individuals are served as reference in addition to the ground truth segmentation. The standard deviation of their F1-scores is 0.7 and the average Pearson correlation coefficient is 0.89. This statistical analysis indicates that our model is robust to the first-frame annotation and can produce reasonable results with inter-variations of reference frame annotations.

Unseen Anatomy Segmentation. To investigate generalisability, we tested our segmentation model on unseen anatomy, *i.e.,* the fetal head. An example result is shown in Fig. 5. Segmenting the fetal head in ultrasound video is challenging due to the significant inter-frame shape changes. Our model still generates valid head segmentation masks over time when given the annotation of the first frame.

3.3 Conclusions

In this paper, we have proposed an automated ultrasound video segmentation method which exploits temporal continuity over video frames. A memory bank is constructed by a memory encoder to extract and store the association between a frame and its segmentation over time. A skip gate is proposed to control the memory module update resulting in an adaptive temporal memory bank for retrieval. Our approach provides new insight using the preceding frames in a memory bank for online video stream segmentation, and it achieves state-of-the-art performance on the placenta and maternal bladder. Experiments on video of an unseen fetal head show the potential of our model to be applied to other ultrasound anatomical segmentation tasks.

Acknowledgments. We acknowledge the ERC (ERC-ADG-2015 694581, project PULSE), the Global Challenges Research Fund (EP/R013853/1, project CALOPUS), EPSRC Programme Grant (EP/T028572/1, project VisualAI), and the InnoHK-funded Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE) Project 2.1 (Cardiovascular risks in early life and fetal echocardiography).

References

1. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 221–230 (2017)
2. Campos, V., Jou, B., Giró-i Nieto, X., Torres, J., Chang, S.F.: Skip RNN: learning to skip state updates in recurrent neural networks. In: International Conference on Learning Representations (2018)
3. Chang, R., Wang, D., Guo, H., Ding, J., Wang, L.: Weakly-supervised ultrasound video segmentation with minimal annotations. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12908, pp. 648–658. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_62
4. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
5. Gleed, A.D., et al.: Automatic image guidance for assessment of placenta location in ultrasound video sweeps. Ultrasound Med. Biol. **49**(1), 106–121 (2023)
6. 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, pp. 770–778 (2016)
7. 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)

8. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring R-CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6409–6418 (2019)
9. Jauniaux, E., Collins, S., Burton, G.J.: Placenta accreta spectrum: pathophysiology and evidence-based anatomy for prenatal ultrasound imaging. Am. J. Obstet. Gynecol. **218**(1), 75–87 (2018)
10. 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)
11. Oppenheimer, L., et al.: Diagnosis and management of placenta previa. J. Obstet. Gynaecol. Can. **29**(3), 261–266 (2007)
12. Qi, H., Collins, S., Noble, A.: Weakly supervised learning of placental ultrasound images with residual networks. In: Valdés Hernández, M., González-Castro, V. (eds.) MIUA 2017. CCIS, vol. 723, pp. 98–108. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60964-5_9
13. 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
14. Schilpzand, M., et al.: Automatic placenta localization from ultrasound imaging in a resource-limited setting using a predefined ultrasound acquisition protocol and deep learning. Ultrasound Med. Biol. **48**(4), 663–674 (2022)
15. Self, A., Gleed, A., Bhatnagar, S., Noble, A., Papageorghiou, A.: Vp18. 01: machine learning applied to the standardised six-step approach for placental localisation in basic obstetric ultrasound. Ultrasound Obstetr. Gynecol. **58**, 172–172 (2021)
16. Self, A., et al.: Developing clinical artificial intelligence for obstetric ultrasound to improve access in underserved regions: protocol for a computer-assisted low-cost point-of-care ultrasound (calopus) study. JMIR Res. Protocols **11**(9), e37374 (2022)
17. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
18. Yang, X., et al.: Towards automated semantic segmentation in prenatal volumetric ultrasound. IEEE Trans. Med. Imaging **38**(1), 180–193 (2018)
19. Zhou, T., Li, L., Bredell, G., Li, J., Unkelbach, J., Konukoglu, E.: Volumetric memory network for interactive medical image segmentation. Med. Image Anal. **83**, 102599 (2023)
20. Zimmer, V.A., et al.: A multi-task approach using positional information for ultrasound placenta segmentation. In: Hu, Y., et al. (eds.) ASMUS/PIPPI -2020. LNCS, vol. 12437, pp. 264–273. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60334-2_26
21. Zimmer, V.A., et al.: Towards whole placenta segmentation at late gestation using multi-view ultrasound images. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11768, pp. 628–636. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32254-0_70