WFUMB Ultrasound Open 2 (2024) 100075

Contents lists available at ScienceDirect

WFUMB Ultrasound Open

journal homepage: www.journals.elsevier.com/wfumb-ultrasound-open

Original contribution

Detection of fetal congenital heart defects on three-vessel view ultrasound videos

Netzahualcoyotl Hernandez-Cruz [a,*], Olga Patey [b], Bojana Salovic [b], Divyanshu Mishra [a], Md Mostafa Kamal Sarker [a], Aris Papageorghiou [b], J. Alison Noble [a]

[a] Institute of Biomedical Engineering, University of Oxford, Old Road Campus Research Building, Oxford, OX3 7DQ, UK
[b] Nuffield Department of Women's & Reproductive Health, University of Oxford, Women's Centre, John Radcliffe Hospital, Oxford, OX3 9DU, UK

ARTICLE INFO

Keywords:
Fetal ultrasound
Congenital heart defects
Segmentation

ABSTRACT

Background: Detecting congenital heart defects (CHDs) is challenging due to the difficulty of identifying subtle abnormalities in fetal heart structures.
Objectives: To develop a deep learning-based method for segmenting vessels in the three-vessel view (3VV) to characterise the vessels by size and spatial relationships to detect abnormal fetal hearts.
Methods: We present a deep learning-based method that takes as input a fetal heart ultrasound (US) video of the three vessels view (3VV) and an anchor frame, which contains the segmentation of the pulmonary artery (PA), aorta (Ao), and superior vena cava (SVC) in the 3VV. The method automatically segments the anatomical structures subsequent to the anchor frame and classifies the US video as normal or abnormal. The method consists of two phases. The first phase combines three residual networks (ResNets) extended with a self-attention block and a refinement module. The second phase extends a ResNet with two CoordConv layers integrating spatial coordinates. We assess segmentation performance using the intersection over union (IoU) and dice similarity coefficient (DSC) metrics and classification of US videos using sensitivity and specificity. We also investigate the tolerance to failure of the method by introducing mislabelled anchor frames. The dataset used in this study consists of 150 US videos of the 3VV; 50 videos were used for training, and 100 videos (50 normal videos, 50 abnormal videos) for testing.
Results: In terms of anatomical structure segmentation accuracy, the method achieves an average IoU of 89.5% (99.5% for PA, 85.0% for Ao, and 84.1% for SVC), and an average DSC of 0.950% (0.946% for PA, 0.969% for Ao, and 0.934% for SVC). Detection of abnormal videos achieved a sensitivity of 0.99 and specificity of 1.0. The tolerance to failure analysis shows a decrease in the sensitivity of 0.023 and 0.015 for normal and abnormal case videos, respectively.
Conclusions: The initial evaluation of our approach to fetal CHDs on 3VV ultrasound videos is promising but requires further refinement and evaluation on a larger dataset to assess clinical utility. The approach is designed to be translatable to low-resource settings where fetal echocardiography experts are unavailable due to the simple acquisition protocol.

1. Introduction

Fetal congenital heart defects (CHDs) are structural cardiac malformations affecting 8 to 9 per 1000 births worldwide [1]. Prenatal detection of CHDs is achieved by fetal echocardiography, a non-invasive technique that uses real-time ultrasound (US) image acquisition to evaluate the fetal heart [2]. In high-income settings with fetal screening programs, assessment of the fetal heart usually takes place in 18-22 week gestational-age fetuses [3], and uses five standard views: situs, four-chamber view (4CH), left ventricular outflow tract view (LVOT), three-vessel view (3VV), and three-vessel trachea view (3VT) [3]. Each view helps to examine intra-cardiac connection and spatial relationship, structural anatomy, chamber and vessel dimensions [4]. Although 60% of CHDs can be detected using the 4CH alone, several major CHDs such as tetralogy of Fallot (the most common defect, afflicting 5% of all CHDs), transposition of the great arteries (the second most common defect afflicting 2% of all CHDs), truncus arteriosus, double-outlet right ventricle and some others can be missed unless views of great arteries are included [5]. During two-dimensional echocardiography, the 3VV

* Corresponding author.
E-mail address: netzahualcoyotl.hernandez-cruz@eng.ox.ac.uk (N. Hernandez-Cruz).

https://doi.org/10.1016/j.wfumbo.2024.100075
Received 31 July 2024; Received in revised form 4 November 2024; Accepted 14 November 2024
Available online 23 November 2024
2949-6683/© 2024 The Authors. Published by Elsevier Inc. on behalf of World Federation for Ultrasound in Medicine and Biology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

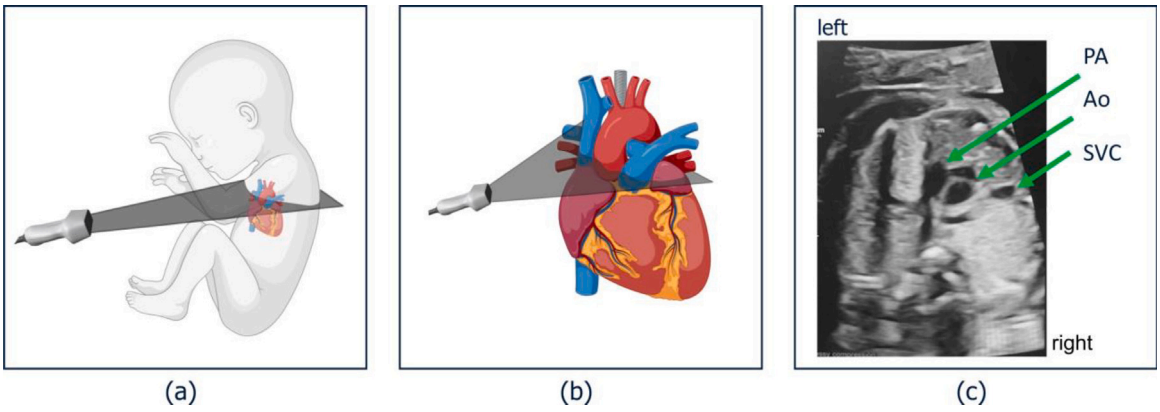

Fig. 1. The 3VV is obtained by moving the transducer in the axial scanning plane across the fetal chest from the 4CH towards the fetal head (cephalic direction). (a) Position of the transducer directed over the fetus. (b) Position of the transducer directed over the fetal heart. (c) 3VV of a normal fetal heart showing the great vessels (pulmonary artery, PA; aorta Ao) and superior vena cava (SVC).

allows the examination of the number, size, position and course of the great vessels (pulmonary artery, PA; aorta Ao) and superior vena cava (SVC) (Fig. 1). Convolutional neural networks (CNNs) can be particularly useful in this context, as they are adept at analysing both spatial and temporal features in US videos, thus aiding accurate assessment of vessel integrity throughout the cardiac cycle.

2. Related work

Good segmentation of fetal heart anatomy is crucial in detecting CHDs as it allows for the accurate delineation and identification of anatomical structures in US scans, ensuring the correct interpretation of potential structural abnormalities. Semantic segmentation is an essential task in medical image analysis. Traditional convolutional neural networks (CNNs), widely used for semantic segmentation, may need help modelling global context due to their local receptive fields. Alternatively, transformers, another type of neural network architecture, have shown excellent performance in capturing long-range dependencies between pixels and learning meaningful image representations [6]. Transformer-based architectures for medical image segmentation include the Vision Transformer (ViT) [7], U-Net++ [8], TransUNet [9], DeiT-Seg [10], among others. These architectures utilise different mechanisms, such as self-attention and positional encoding, to capture global context and to improve the segmentation performance. ViT is a purely transformer-based architecture that uses self-attention mechanisms to capture global dependencies between pixels in an image; it treats each image as a sequence of tokens and then feeds them to multiple transformer layers to operate [7]. U-Net++ and TransUNet are modified versions of the U-Net [11] architecture that integrates transformers into the decoder path. They use a combination of skip connections and transformers to improve feature propagation and context integration [8]. Similarly, DeiT-Seg offers a variant of DeiT vision transformer architecture adapted for image segmentation tasks, combining transformers with traditional CNNs to capture global and local information [10].

In addition to the above, hierarchical and pyramidal transformer-based architectures such as Swin [12], CvT [13], CoaT [14], LeViT [15], and Twins [16] have focused on enhancing the local continuity of features and removing fixed-size position embedding to improve the performance of transformers, demonstrating the potential of a pure transformer backbone compared to CNN counterparts in segmentation tasks. While several architectures have been proposed, these Transformer-based methods have very low efficiency in terms of computational and memory resources and, thus, are challenging to deploy in real-time applications. Lightweight architectures include the Space-Time Memory network (STM) [17], Recurrent Network for Video Object Segmentation (RVOS) [18], Space-Time Convolutional Network with improved memory coverage (STCN) [19], and Long-Term Video Object Segmentation with an Atkinson–Shiffrin Memory Model (XMem) [20]. STM [17] uses a memory-augmented neural network where features from previous frames are stored and retrieved for accurate mask propagation. This approach addresses occlusions and appearance changes over time, improving segmentation accuracy and temporal coherence. RVOS [18] employs a recurrent neural network to model temporal dependencies. It integrates a convolutional LSTM to propagate object segmentation masks through time, effectively capturing long-term temporal information. The recurrent structure allows the network to handle significant appearance changes and occlusions, thereby enhancing the robustness of the segmentation results over extended video sequences. STCN [19] utilises a long-term memory of past frames to inform the segmentation of the current frame. XMem [20] includes components for sensory storage, short-term memory, and long-term memory. This hierarchical memory structure balances computational efficiency and segmentation accuracy over long video sequences, achieved by strategically leveraging different memory components to handle varying temporal spans.

Research on the automatic segmentation of the 3VV is limited. Some studies [21–23] use knowledge distillation, where a compact model (student) is trained under the guidance of a larger model (teacher). For instance, in [21], channel-wise knowledge distillation is chosen over pixel-wise distillation due to its effectiveness in aligning point-wise classification scores [24]; this method converts feature maps on each channel to probability maps, aligning the channel-wise probabilities of the teacher and student models with a divergence-based loss. Other studies [22] adopted the DeepLabv3+ [25] architecture with channel-wise distillation to segment three vessels in fetal heart US images, where both teacher and student networks share the same architecture, with training data involving cropped regions of interest (RoI) from full-size images for teacher model training. The teacher network's logit output from cropped inputs distils knowledge into the student model, which is then trained using full-size images. Another study, [23], proposes a multi-task learning method for fetal heart diagnosis, covering segmentation, classification, and detection; this approach uses the Mask-RCNN architecture, which includes region proposal networks for feature extraction and CNN-based models for classification, detection, and segmentation.

To our knowledge, our paper described the first work on segmenting vessels in the 3VV over US videos. We present a method that operates in two phases: first, it segments frames in the US videos, and second, it characterises the vessels by size and spatial relationships to detect abnormal fetal hearts. We validate the method with experiments involving 150 US videos.

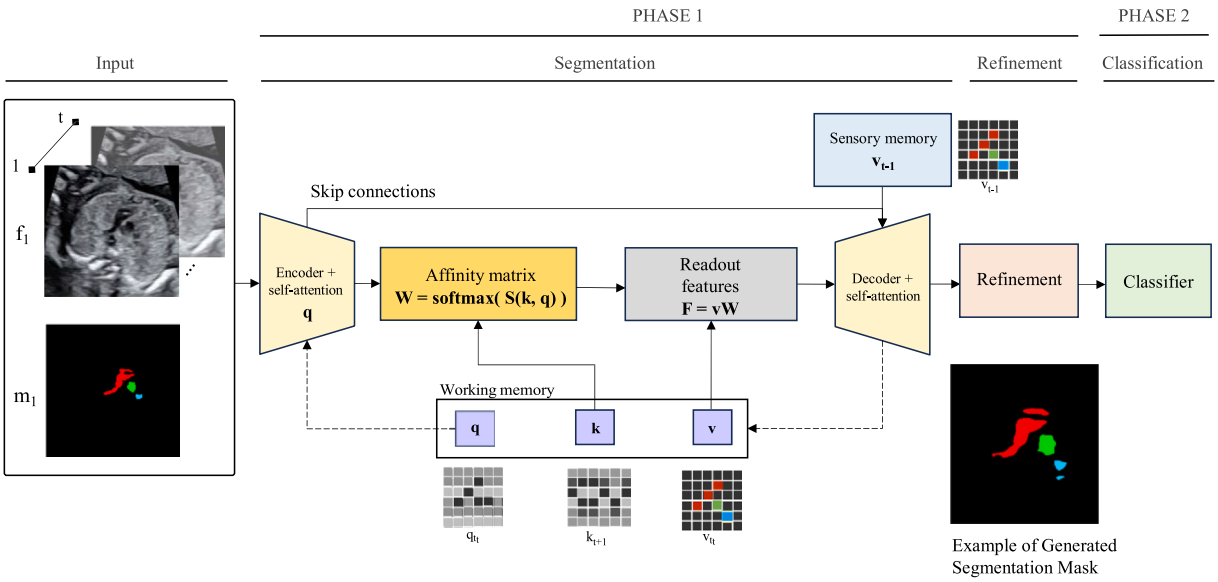

Fig. 2. The method architecture consists of two phases. The first phase extracts q and k, corresponding to the fetal heart anatomy embeddings in the anchor frame. W and S are then calculated based on the pairwise similarity between q and k followed by a SoftMax operation. Readout features are obtained by mapping v incorporating information from both the image and the segmentation to W; embeddings are then passed to a decoder to generate a segmentation, followed by their refinement. The second phase uses a CoordConv network to detect abnormal US videos.

3. Methods

The method follows the one-shot approach [26], where the first frame of the video (anchor frame) contains the segmentation of the fetal heart anatomical structures in the 3VV. The work described by [20] informs the method, enhancing its architecture with self-attention blocks and a refinement component for segmenting the 3VV. Additionally, we incorporate a CoordConv [27] network to classify US videos as normal or abnormal by analysing the size, alignment, and number of vessels.

Fig. 2 shows the method receives a US video of the 3VV and an anchor frame. The method characterises the segmented anatomy the anchor frame provides and propagates its segmentation across subsequent frames. The method refines the outputted segmentations and classifies the US video as normal or abnormal.

The method comprises two memory modules (working memory and sensory memory), three CNNs extended with self-attention blocks (query encoder, decoder, and classifier), and three image analysis components (affinity matrix, readout features, and refinement); further information on memory, affinity matrix, and readout features can be found at [20]. In the following section, we describe the refinement components and CNN architecture (Encoder, Decoder, and Classifier), which are the original contributions of our method.

The method initialises the memory modules with random values. For subsequent iterations, the method utilises the working memory. The encoder extracts query-specific image features, the decoder uses the output of the working memory step to generate the segmentations, and the classifier further categorises the generated segmentations for abnormal video classification. Considering the input segmentation, the affinity matrix and readout feature components characterise relevant features from the frames in the sweep. Refinement focuses on filling gaps in the generated segmentation to address discontinuities in segmented regions.

3.1. Refinement component

The refinement component consists of two algorithms Sklansky's [28] employed to simplify the contours of m_t by removing redundant points along the edges of the segmentation. The simplified contours are then used to find the RoI representing the gap between two blobs defined by the nearest sequence of points, and the Douglas-Peucker [29] used for the simplification of polygonal curves (represented by a sequence of points along the contours).

3.2. Encoder and decoder

The encoder and decoder use a ResNet50 architecture as the backbone. We extend the backbone using self-attention blocks to capture dependencies and relationships within input sequences of features [30]. The self-attention blocks are placed at the beginning of each residual block and consist of two convolutional layers followed by Sigmoid activation. The query encoder component generates essential feature information for feature extraction to create the query (q). The decoder component reconstructs the output based on the encoded information. It involves concatenating the hidden representation (h_{t-1}) with the readout feature (F), then iteratively upsampling them while incorporating skip connections from the query encoder at multiple stages. This process refines and reconstructs the final output, culminating in a single-channel logit generated through a convolution operation. Finally, this logit is bilinearly upsampled to match the input resolution.

3.3. Classifier

Similarly to the encoder and decoder, the classifier uses a ResNet50 and self-attention blocks and adds a CoordConv [27] layer at the network's end. The CoordConv layer adds two extra channels (i and j coordinates) to the associated convolutional layers, enabling spatial information at each pixel to bolster its ability to discern and leverage spatial relationships in the segmentations. This is necessary for the classifier to characterise the properties of the vessels in terms of size, alignment, and number of vessels. Our method follows clinical practice diagnostic criteria [31], which indicate that a normal 3VV consists of three visible vessels (PA, Ao, SVC), their size follows a proportion such as the PA ≥ Ao and Ao > SVC, and they are aligned one after the other (the PA followed by Ao and the Ao followed by SVC). Our method declares any example that fails the criteria as abnormal 3VV. For example, in Fig. 3, we illustrate a normal and an abnormal case (diagnosed with coarctation of the aorta) where there was a disproportion between great vessels (Ao and PA). Our method currently distinguishes (by binary classification) normal and abnormal videos, whereas the classification of specific abnormalities may be addressed in future work.

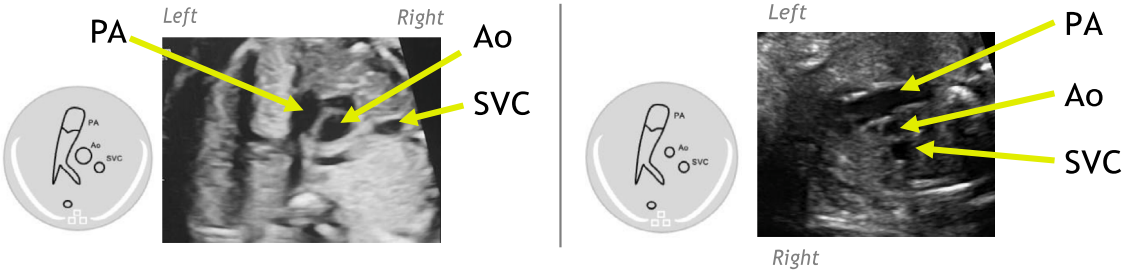

Fig. 3. Schematics and US images showing normal and abnormal 3VV examples. On the left, the schematic and image show a normal heart consisting of three vessels, normal proportion in size (PA ≥ Ao and Ao > SVC), and alignment (the PA on the left side of the fetal chest followed by Ao and then SVC on the right side of the fetal chest). On the right, the schematic and image show an abnormal heart (diagnosed with coarctation of the aorta), which is different by having a disproportion of the great vessels (Ao < PA).

3.4. Implementation

Model training was performed using a batch size of 8 on two Nvidia RTX 6000 GPUs with 24 GB of VRAM and 16 CPU cores (see Appendix C for minimal technical requirements for implementation). We used bootstrapped cross-entropy loss and dice loss, assigning equal weight. AdamW [32] optimisation was utilised with a learning rate set at $1e-5$ and a weight decay of 0.05. For the segmentation tasks, we used intersection over union (IoU) and dice similarity coefficient (DSC) metrics. We used sensitivity and specificity to assess the detection of abnormal heart US videos.

To compare the performance of our method against the state-of-the-art, we implemented the following methods accordingly to the original papers: Space-Time Memory network (STM) [17], Long-Term Video Object Segmentation with an Atkinson–Shiffrin Memory Model (XMem) [20], Recurrent Network for Video Object Segmentation (RVOS) [18], and Space-Time Convolutional Network with improved memory coverage (STCN) [19]. The pixel resolution of the input data was adjusted to meet the architectural requirements of each network: STM (224 × 224), XMem (512 × 512), RVOS (448 × 448), STCN (224 × 224), and our method (512 × 512).

3.5. Dataset

The data used in this paper comes from the study Clinical Artificial Intelligence Models in Fetal Echocardiography (CAIFE). The CAIFE study data consisted of US videos on the fetal hearts and included pregnant women > 18 years of age in their second trimester (13–27 weeks; mean gestational age of 20 weeks). Ethics approval was granted by the East Midlands - Leicester Central Research Ethics Committee (REC Reference 23/EM/0023).

Data was obtained by experienced fetal cardiologists using standard curvilinear transducers and different US machines. The US machine models consisted of General Electric Voluson (E8 and E10), Fujifilm (ARIETTA 850), Samsung (HERA-W10), Sonoscape (P60), and Canon (TUS-AI800), using their preferred imagine tint map. The videos were extracted on Digital Imaging and Communications in Medicine (DICOM[1]) format using ViewPoint v5 as the image management system.

Videos were extracted from the DICOM files at the frame level to ensure high-quality graphics. Each frame was converted into Tagged Image File Format (TIFF) and converted into greyscale using DCMTK (DICOM Toolkit)[2] and then anonymised using ImageMagic[3] library. The image resolution was preserved as available in the original DICOM files. The pixel resolution varied between videos, for example, 1024 × 768 and 1280 × 960.

Anonymisation involved placing a 65 pixels-high black rectangle at the header section of each image, spanning the full width to cover participant, hospital, and scan date details. Additionally, identifiable information (such as participant, operator, and hospital details) was removed from the DICOM metadata using the dcmdump[4] library. Anonymised images were converted into MP4 videos (using the H.264 codec) via OpenCV[5] library. The video's approximate duration is 1.5 s (60 frames per video).

For the current work, two datasets of US videos (one per participant) were selected. Dataset-A consists of 50 normal US videos where all frames are manually annotated. Dataset-B consists of 100 videos (50 normal and 50 abnormal) where, in each case, only the anchor frame is manually annotated. The combined number is 150 US videos (7193 frames in total). The US videos with abnormal cases included the following CHD: hypoplastic left heart syndrome (20 videos), tetralogy of Fallot with left aortic arch (10 videos), complete transposition of the great arteries with intact interventricular septum (10 videos), and coarctation of the aorta (10 videos). The selection criteria included videos of the 3VV with clearly visible anatomy of the pulmonary artery (PA), aorta (Ao), and superior vena cava (SVC) and diagnosis as either normal or abnormal (positive for a CHD).

Appendix B provides a breakdown of the participants' pathology and the scanning system. Each participant's identifier consists of six characters: the first character represents the participant, the second indicates pathology (with 'N' for normal and 'A' for abnormal), and the third is followed by four arbitrary digits.

4. Experiments and results

We conducted two experiments: first, to select a segmentation architecture suitable for our task with validation on normal US videos; second, to evaluate the method for detecting normal and abnormal videos and to investigate the method's tolerance to failure by mislabelling the anchor frame.

4.1. Experiment 1

The first experiment assessed the segmentation of the vessels in the 3VV. We used Dataset-A, dividing it into three subsets (80% training, 10% validation, and 10% testing) to investigate the performance of our referenced CNN-based video processing methods: STM, XMem, RVOS, STCN, and our method. Each network model was trained using the training subset to produce the probability scores for each substructure (PA, Ao, and SVC). As shown in Table 1, our method achieves a mean IoU of 0.895 (0.995 for PA, 0.850 for Ao, and 0.841 for SVC) and mean DSC of 0.950 (0.946 for PA, 0.969 for Ao, and 0.934 for SVC) at 31.2 frames per second (FPS), which is superior to STCN in IoU by 0.071 (0.020 for PA and 0.109 for Ao) and in DSC by 0.036 (0.010 for PA,

[1] https://www.dicomstandard.org/current/
[2] https://dicom.offis.de/en/
[3] https://imagemagick.org/
[4] https://manpages.ubuntu.com/manpages/focal/man1/dcmdump.1.html
[5] https://opencv.org/

Table 1
Comparison of four state-of-the-art methods for one-shot video segmentation and our method. Our method achieves the highest intersection over IoU and DSC.

Method	IoU | DSC (PA)	IoU | DSC (Ao)	IoU | DSC (SVC)	Mean IoU | DSC	FPS
STM	0.378 | 0.047	0.022 | 0.048	0.019 | 0.046	0.140 | 0.047	10.6
XMem	0.855 | 0.706	0.580 | 0.723	0.539 | 0.697	0.658 | 0.709	26.6
RVOS	0.962 | 0.792	0.648 | 0.811	0.651 | 0.782	0.754 | 0.795	22.7
STCN	0.975 | 0.936	0.756 | 0.932	0.741 | 0.873	0.824 | 0.914	24.4
Our method	**0.995** | **0.946**	**0.850** | **0.969**	**0.841** | **0.934**	**0.895** | **0.950**	**31.2**

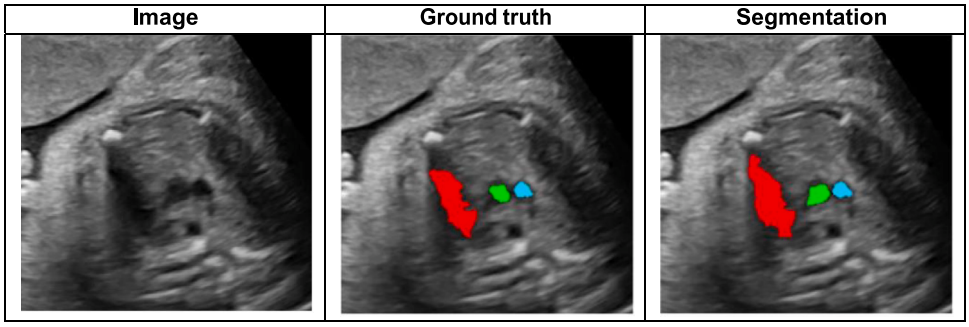

Fig. 4. Qualitative results for the segmentation of 3VV compared to the manually derived GT. These images have been manually overlapped (bottom) and 50% zoomed in to contrast differences for illustrative purposes. The red, green, and blue colour codes represent the PA, Ao, and SVC, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
For brevity, we provide a breakdown sample of performance case-based for 30 of the subset of tested US videos. The first column shows the list of participants, followed by classification sensitivity and specificity. The result of the performance from the videos not listed is 1.0.

Participants	Sensitivity	Specificity
*PN*0191, *PN*0195, *PN*0197, *PN*0198, *PN*0199, *PN*0202, *PN*0203, *PN*0204, *PN*0205, *PN*0206, *PN*0207, *PN*0208, *PN*0213, *PA*0458, *PA*0516, *PA*0769, *PA*0812, *PA*0900, *PA*0962, *PA*0975, *PA*1057, *PA*1133, *PA*1150, *PA*1377, *PA*1485, *PA*1509, *PA*1601, *PA*1624	1.0	1.0
*PN*0214	0.85	1.0
*PN*0193	0.96	1.0
Average	**0.99**	**1.0**
SD	**0.02**	**0.0**

0.037 for Ao, and 0.061 for SVC) and 4.6 FPS compared to XMem. Qualitative evaluation confirms that the method's segmentation is more precise than other methods in segmenting, such as the valve location and underlying vessels.

Fig. 4 presents qualitative results from a challenging segmentation case. This image contains a shadow from the fetal rib running along the length of the pulmonary artery. Such a case may be difficult to segment for a human due to the poor border definition of vessels and their structures. The correct vessel size is important in clinical practice, as a disproportion between the great vessels is associated with several congenital heart defects. For this example, our model overestimates the dimension of the pulmonary artery compared to the manually determined ground truth (GT). An explanation for this difference between the two methods is that the pulmonary artery is acoustically shadowed, giving an impression that the PA is larger than it is physically. See Appendix A for a presentation of further qualitative results.

4.2. Experiment 2

This experiment aimed to classify US videos as containing normal or abnormal fetal anatomy. We performed US video classification using the trained model, which reached the highest mean IoU and DSC in experiment 1 and used Dataset-B as a testing dataset only. We also investigated the method's tolerance to failure by deliberately mislabelling the anchor frames.

Table 3
Sensitivity results for the introduction of mislabelled anchor frames at different positions, where 'No Frames' stands for no mislabelled frames; 'First frame' stands for mislabelling the first frame of the video; 'Half Quartile' stands for mislabelling the frame at position 12.5% of the video; 'Quartile' stands for mislabelling the frame at position 25% of the video.

	No Frame	First Frame	Half Quartile	Quartile	Diff
Normal	0.996	0.939	0.942	0.940	0.023
SD	0.059	0.079	0.093	0.085	0.026
Abnormal	1.000	0.975	0.981	0.985	0.015
SD	0.000	0.031	0.027	0.025	0.025

4.2.1. Detection of abnormal US videos

Overall evaluation achieved a sensitivity of 0.99 (SD of 0.02) for normal videos and 1.0 (SD of 0.0) for abnormal videos, and specificity of 1.0 for both classes. Table 2 shows the breakdown of the video assessment.

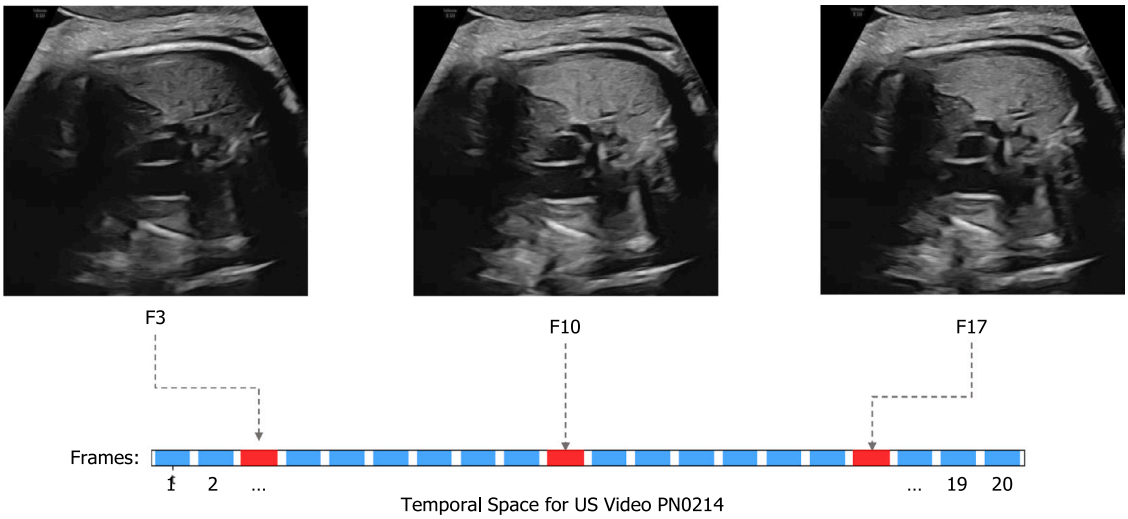

Fig. 5. Mislabelled frames in *PN*0214 (normal heart). Correct frames are represented in blue, and mislabelled frames are in red. US of the mislabelled frames ($F3$, $F10$, and $F17$) are shown above the timeline. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

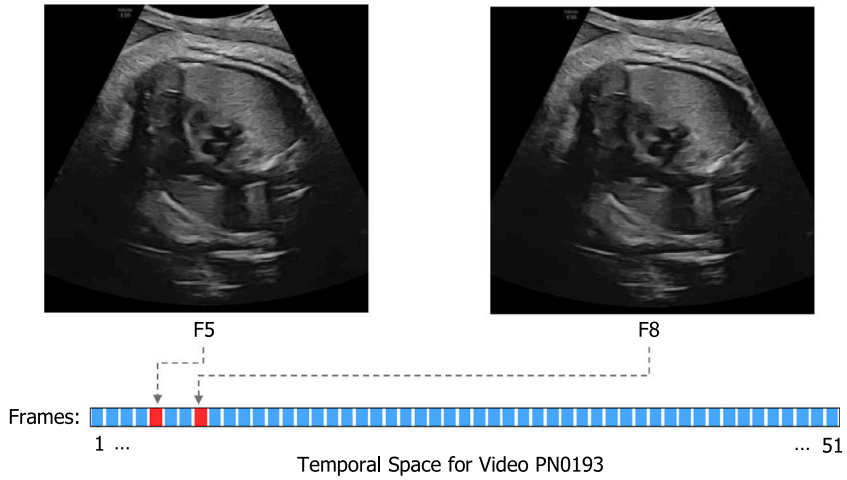

Fig. 6. Mislabelled frames in *PN*0193 (normal heart). Correct frames are represented in blue, and mislabelled frames are in red. A sample of the mislabelled US frames ($F5$ and $F8$) is shown above the timeline. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In the normal video *PN*0214, three frames were incorrectly labelled as abnormal instead of normal; this happened due to acoustic shadows generated from the ribs running across all images; Fig. 5 locates the mislabelled frames ($F3$, $F10$, and $F17$) in the temporal space representation of the US scan video. This case is challenging even for a trained sonographer. In *PN*0193, two frames were mislabelled; in these frames, the aorta (Ao) appears smaller than the pulmonary artery (PA), which corresponds to an abnormal heart; a clinical expert (OP) confirmed that this occurred due to the fetal motion during scanning. Fig. 6 locates the mislabelled frames ($F5$, and $F8$) in the temporal space representation of the US scan video.

4.2.2. Tolerance to failure

The method follows the one-shot approach [26], where the first frame of the video (anchor frame) is paired with a manual segmentation of the fetal heart anatomical structures in that frame. In this experiment, we investigate the tolerance to failure of the method by mislabelling the anchor frame. Mislabelled anchor frames were defined by labelling extra vessels where normal US videos would appear abnormal, and abnormal US videos would appear normal. Mislabelled frames were placed at different temporal positions in the video: the first frame, the frame at position 12.5% of the video (half quartile), and the frame at position 25% of the video (quartile). We used the 30 US videos from Dataset-B (15 normal videos and 15 abnormal videos) for evaluation.

Table 3 shows the results for normal and abnormal US videos when anchor frames are mislabelled. We observed that when the anchor frames mislabelled anatomical structures like the ventricles (in addition to the expected three vessels in the 3VV; PA, Ao, and SVC), method sensitivity decreases by an average of 0.023 (SD 0.026) and 0.015 (SD 0.025) for normal and abnormal US videos respectively. We noted that the highest decrement occurred when the anchor frame at the diastole phase of the cardiac cycle was mislabelled (see Fig. 7).

In this assessment of tolerance to failure, we observed that the classification sensitivity varies across the four variations in Table 3 (no frame, first frame, half quartile, and quartile) because the morphological appearance of heart anatomical structures, like ventricles, changes over the fetal cardiac cycle. Fig. 7 illustrates the changes in ventricle appearance.

5. Conclusion

We developed a method to segment the fetal heart anatomy in the 3VV and classify real-world US videos of the fetal heart as normal or abnormal. We note two limitations of the current approach. First,

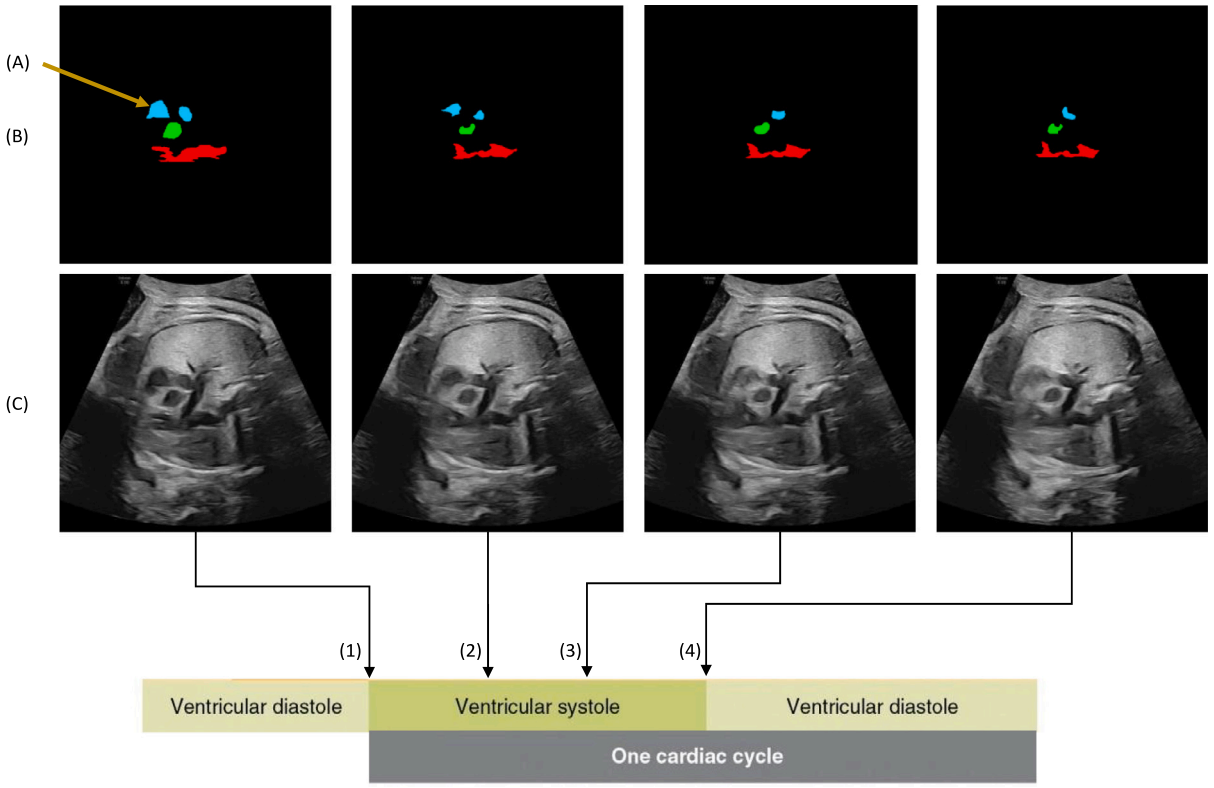

Fig. 7. Image (A) shows the morphological changes of a mislabelled heart vessel. Images in (B) show the segmentations generated by the method followed by their respective US image in (C). The leftmost column shows the smallest mislabelled vessel during the (1) end-diastolic phase of the cardiac cycle — when the heart's ventricles are filled with blood just before they contract. (2, 3) Shows the changes along the transition to the systole phase of the cardiac cycle. The right-most images (4) show the mislabelled heart vessels are the biggest size at the end-systolic cardiac phase when the ventricles have contracted and ejected the blood.

the method assumes the existence of an anchor frame to segment the vessels in the US video; second, the data used in the study consisted of 50 abnormal case videos, a restriction imposed by data availability at the time of doing this work. Although this is a good number for an initial study, a larger abnormal dataset is needed to capture expected variability in abnormal case videos. In future work, we plan to extend our method to automatically generate the anchor frame segmentations and report results on more abnormal case US videos. In principle, our method is translatable to low-resource settings where fetal echocardiography experts are unavailable due to the simple acquisition protocol. In the future, this could aid in adopting clinical US pregnancy assessments in areas where trained sonographers are scarce or non-existent.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: J. Alison Noble reports financial support was provided by InnoHK-funded Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE). J. Alison Noble reports financial support was provided by Oxford Biomedical Research. J. Alison Noble reports financial support was provided by European Research Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by grants from the InnoHK-funded Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE), the Oxford Biomedical Research Centre, the European Research Council ERC-ADG-2015 694581, the UKRI (UK Research and Innovation) grant reference EP/X040186/1 (Turing AI Fellowship: Ultra Sound Multi-Modal Video-based Human-Machine Collaboration), and the EPSRC (Engineering and Physical Research Council) Programme Grant EP/T028572/1, VisualAI.

Appendix A. Comparison of methods qualitative results

Table 4 shows a sample of qualitative results of our method segmentation compared to the ground truth and state-of-the-art methods for four participants not seen by the model. Qualitative evaluation confirms that our method's segmentation is more precise than the other methods in showing the valve location and underlying vessels. Our method's segmentations align with ground truth, demonstrating its reliability in delineating vessel boundaries across the different frames from the video clips.

Appendix B. Breakdown of participants and US systems used

Tables 5–8 show a breakdown of participants' US machines used during scanning and pathologies. Each participant's identifier consists of six characters: the first character represents the participant, the second indicates pathology ('N' for normal and 'A' for abnormal), and the third is followed by four arbitrary digits. US machine models included General Electric Voluson (E8, E10, and S10), Fujifilm (ARIETTA 850), Samsung (HERA-W10), Sonoscape (P60), and Canon (TUS-AI800). Pathologies included hypoplastic left heart syndrome (HLHS), tetralogy of Fallot with left aortic arch (TOF-LAA), complete transposition of the great arteries with the intact interventricular septum (TOF-IVS), and coarctation of the aorta (COA).

Table 4
Qualitative results. The original fetal heart ultrasound frame is shown in the first row, followed by the ground truth (GT) and the segmentation results from the state-of-the-art methods.

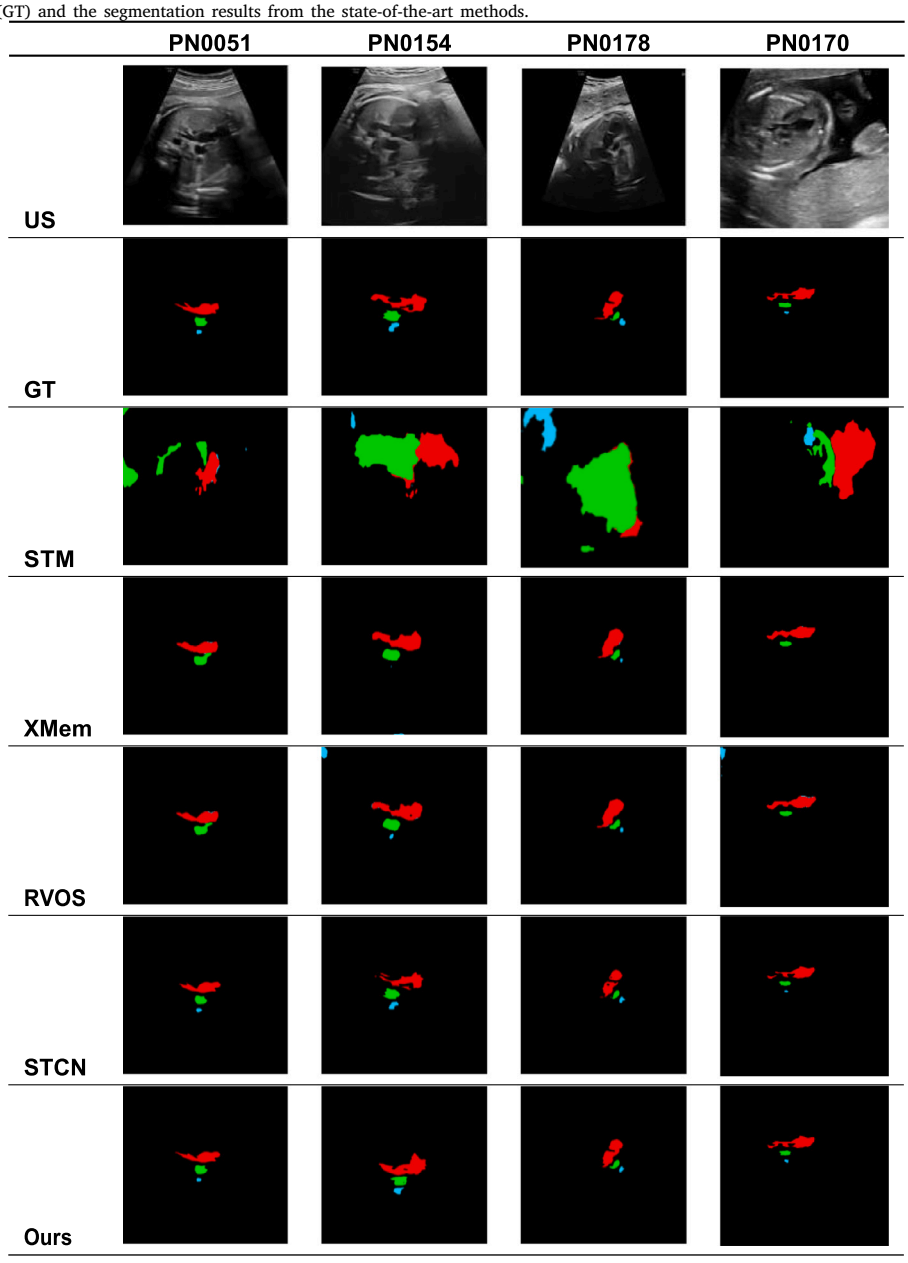

Table 5
This table outlines participants and the ultrasound system used for scanning, corresponding to normal US video diagnoses.

Participant	Diagnosis	Machine Manufacturer	Machine Model	Machine Serial Number	Software Version
PN0004	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0005	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0008	Normal	General Electric	Voluson E8	E39416	18.x.x VE
PN0021	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0022	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0023	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0024	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0027	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0030	Normal	General Electric	Voluson E8	E39416	18.x.x VE
PN0031	Normal	General Electric	Voluson E8	E39416	18.x.x VE
PN0039	Normal	General Electric	Voluson E8	E39416	18.x.x VE
PN0042	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0058	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0059	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0071	Normal	General Electric	Voluson E8	E39416	18.x.x VE

(*continued on next page*)

Table 5 (*continued*).

Participant	Diagnosis	Machine Manufacturer	Machine Model	Machine Serial Number	Software Version
PN0078	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0079	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0083	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0091	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0092	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0095	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0110	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0129	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0135	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0140	Normal	General Electric	Voluson E10	E82066	21.x.x VE
PN0141	Normal	General Electric	Voluson E10	E82066	21.x.x VE
PN0142	Normal	General Electric	Voluson E10	E82066	21.x.x VE
PN0143	Normal	General Electric	Voluson E10	E82066	21.x.x VE
PN0145	Normal	General Electric	Voluson E10	E82066	21.x.x VE
PN0159	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0175	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0176	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0177	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0183	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0192	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0206	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0207	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0208	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0213	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0214	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0191	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0193	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0195	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0197	Normal	General Electric	Voluson E10	E82066	21.x.x VE
PN0198	Normal	General Electric	Voluson E10	E82066	21.x.x VE
PN0198	Normal	General Electric	Voluson E10	E82066	21.x.x VE
PN0202	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0203	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0204	Normal	General Electric	Voluson E10	E81395	21.x.x VE
PN0205	Normal	General Electric	Voluson E10	E81395	21.x.x VE

Table 6
This table outlines participants and the ultrasound system used for scanning, corresponding to hypoplastic left heart syndrome (HLHS) US video diagnoses.

Participant	Diagnosis	Machine Manufacturer	Machine Model	Machine Serial Number	Software Version
PA0403	HLHS	Fujifilm	ARIETTA 850	NA	20200530
PA0458	HLHS	Fujifilm	ARIETTA 850	NA	20200530
PA0516	HLHS	Fujifilm	ARIETTA 850	NA	20200530
PA0525	HLHS	Fujifilm	ARIETTA 850	NA	20200530
PA0769	HLHS	Canon	TUS-AI800	5LB2082031	V6.5 SP0004*
PA0868	HLHS	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA0869	HLHS	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA0900	HLHS	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA0962	HLHS	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA0975	HLHS	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1063	HLHS	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1101	HLHS	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1176	HLHS	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1359	HLHS	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1370	HLHS	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1469	HLHS	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1638	HLHS	Canon	TUS-AI800	AED1722087	V2.3 SP0000*
PA1639	HLHS	Canon	TUS-AI800	AED1722087	V2.3 SP0000*
PA1693	HLHS	Canon	TUS-AI800	AED1722087	V2.3 SP0000*
PA1718	HLHS	Canon	TUS-AI800	AED1722087	V2.3 SP0000*

Appendix C. Models technical specifications

We used different models to investigate the performance of CNN-based video processing, including Space-Time Memory network (STM), Long-Term Video Object Segmentation with an Atkinson–Shiffrin Memory Model (XMem), Recurrent Network for Video Object Segmentation (RVOS), Space-Time Convolutional Network with improved memory coverage (STCN), and our method. In Table 9, we provide computational requirements specifically for implementing (inference), including the number of parameters, size, and minimum resources; these resources encompass GPU memory, CPU, and RAM, as detailed below:

- GPU Memory (VRAM): A general rule of thumb is to have enough VRAM to accommodate the model size and input data, with a buffer for intermediate activations. This can often be 1.2 to 1.5 times the model size.

Table 7
This table outlines participants and the ultrasound system used for scanning, corresponding to tetralogy of Fallot with left aortic arch (TOF-LAA) and complete transposition of the great arteries with the intact interventricular septum (TOF-IVS).

Participant	Diagnosis	Machine Manufacturer	Machine Model	Machine Serial Number	Software Version
PA0273	TOF-LAA	Fujifilm	ARIETTA 850	NA	20200530
PA0273	TOF-LAA	Fujifilm	ARIETTA 850	NA	20200530
PA0500	TGA-IVS	Fujifilm	ARIETTA 850	NA	20200530
PA0521	TGA-IVS	Fujifilm	ARIETTA 850	NA	20200530
PA0703	TGA-IVS	General Electric	Voluson E10	E66394	17.x.x VE
PA0739	TGA-IVS	Fujifilm	ARIETTA 850	NA	20200530
PA0799	TOF-LAA	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA0812	TOF-LAA	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA0837	TGA-IVS	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA0942	TOF-LAA	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1057	TGA-IVS	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1061	TOF-LAA	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1377	TOF-LAA	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1485	TGA-IVS	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1509	TGA-IVS	Canon	TUS-AI800	AED1722087	V2.3 SP0000*
PA1577	TGA-IVS	Canon	TUS-AI800	AED1722087	V2.3 SP0000*
PA1590	TGA-IVS	Canon	TUS-AI800	AED1722087	V2.3 SP0000*
PA1601	TOF-LAA	Canon	TUS-AI800	AED1722087	V2.3 SP0000*
PA1603	TOF-LAA	Canon	TUS-AI800	AED1722087	V2.3 SP0000*
PA1801	TOF-LAA	Fujifilm	ARIETTA 850	NA	20200530

Table 8
This table outlines participants and the ultrasound system used for scanning, corresponding to coarctation of the aorta (COA).

Participant	Diagnosis	Machine Manufacturer	Machine Model	Machine Serial Number	Software Version
PA0322	COA	Fujifilm	ARIETTA 850	NA	20200530
PA0499	COA	Fujifilm	ARIETTA 850	NA	20200530
PA0767	COA	Samsung	HERA W10	BS100AHKB	NA
PA0967	COA	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1133	COA	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1150	COA	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1223	COA	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1423	COA	Canon	TUS-AI800	AED1722087	V2.3 SP0102*
PA1425	COA	SonoScape	P60	NA	NA
PA1624	COA	Canon	TUS-AI800	AED1722087	V2.3 SP0000*

Table 9
Technical specifications of CNN-based video processing methods. The first column lists the input size in pixels, the second column presents the number of parameters in millions, and the third column presents the size of the parameters in megabytes. The rightmost column estimates the minimum computational requirements needed for implementation.

Method	Input (px)	Parameters (M)	Size (MB)	Requirements
STM	224 × 224	39	148	GPU: 3-4 GB, CPU: 4 cores, RAM: 4-8 GB
XMem	512 × 512	62	237	GPU: 4-6 GB, CPU: 4-6 cores, RAM: 8-12 GB
RVOS	448 × 448	30	114	GPU: 3-4 GB, CPU: 4 cores, RAM: 4-8 GB
STCN	224 × 224	54	208	GPU: 4-6 GB, CPU: 4-6 cores, RAM: 8-12 GB
Our method	512 × 512	74	285	GPU: 6-8 GB, CPU: 4-8 cores, RAM: 12-16 GB

- CPU: A multi-core CPU is generally recommended. At least 4-8 cores with a clock speed of 3.0 GHz or higher are advisable for handling preprocessing and data loading.
- System RAM: The system RAM should also be sufficient to handle the processed data. A good estimate would be 1.2 to 1.5 times the model's size.

References

[1] Dolk H, Loane M, Garne E. Congenital heart defects in Europe: prevalence and perinatal mortality. Circulation 2011;123:841–9.

[2] Donofrio MT, Moon-Grady AJ, Hornberger LK, Copel JA, Sklansky MS, Abuhamad A, et al. Diagnosis and treatment of fetal cardiac disease: a scientific statement from the American heart association. Circulation 2014;129:2183–242.

[3] Carvalho JS, Axt-Fliedner R, Chaoui R, Copel JA, Cuneo BF, Goff D, et al. ISUOG practice guidelines. Ultrasound Obstet Gynecol 2023;61:788–803.

[4] Burns J, Basken A, Acosta R, Garnier-Villarreal M, Kulkarni A, Hayes DA. Identification of research priorities in CHD: empowering patients and families through participation in the development of formal research agendas. Cardiol Young 2022;1–6.

[5] Huml M, Fremuth J, Jehlička P. Cyanotic heart disease. Cesko-Slovenska Pediatrie 2023;78:7–14.

[6] Fiorentino MC, Villani FP, Cosmo MD, Frontoni E, Moccia S. A review on deep-learning algorithms for fetal ultrasound-image analysis. Med Image Anal 2022;83.

[7] Yuan L, Chen Y, Wang T, Yu W, Shi Y, Jiang Z-H, et al. Tokens-to-token ViT: Training vision transformers from scratch on ImageNet. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 558–67.

[8] Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. UNet++: A nested U-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Cham: Springer International Publishing; 2018, p. 3–11.

[9] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: Transformers make strong encoders for medical image segmentation. 2021, arXiv:2102.04306.

[10] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jegou H. Training data-efficient image transformers and distillation through attention. In: Meila M, Zhang T, editors. Proceedings of the 38th international conference on machine learning. Proceedings of machine learning research, vol. 139, PMLR; 2021, p. 10347–57.
[11] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical image computing and computer-assisted intervention – MICCAI 2015. Cham: Springer International Publishing; 2015, p. 234–41.
[12] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 10012–22.
[13] Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, et al. CvT: Introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 22–31.
[14] Xu W, Xu Y, Chang T, Tu Z. Co-scale conv-attentional image transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 9981–90.
[15] Graham B, El-Nouby A, Touvron H, Stock P, Joulin A, Jégou H, et al. Levit: A vision transformer in ConvNet's clothing for faster inference. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 12259–69.
[16] Chu X, Tian Z, Wang Y, Zhang B, Ren H, Wei X, et al. Twins: Revisiting the design of spatial attention in vision transformers. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang P, Vaughan JW, editors. Advances in neural information processing systems,Vol. 34. Curran Associates, Inc.; 2021, p. 9355–66.
[17] Oh SW, Lee J-Y, Xu N, Kim SJ. Space-time memory networks for video object segmentation with user guidance. IEEE Trans Patterns Anal Mach Intell 2019.
[18] Ventura C, Bellver M, Girbau A, Salvador A, Marques F, Giro-i Nieto X. RVOS: End-to-end recurrent network for video object segmentation. In: Conference on computer vision and pattern recognition. 2019, p. 5277–86.
[19] Cheng HK, Tai Y-W, Tang C-K. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang P, Vaughan JW, editors. Advances in neural information processing systems,Vol. 34. Curran Associates, Inc.; 2021, p. 11781–94.
[20] Cheng HK, Schwing AG. XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model. In: Lecture notes in computer science (lecture notes in artificial intelligence and lecture notes in bioinformatics), vol. 13688 LNCS, Springer Science and Business Media Deutschland GmbH; 2022, p. 640–58.
[21] Han CS, Lee KM. Channel-wise attention and channel combination for knowledge distillation. In: Proceedings of the international conference on research in adaptive and convergent systems. New York, NY, USA: Association for Computing Machinery; 2020, p. 72–6.
[22] Cai Q, Chen R, Li L, Huang C, Pang H, Tian Y, et al. The application of knowledge distillation toward fine-grained segmentation for three-vessel view of fetal heart ultrasound images. Comput Intell Neurosci 2022;2022.
[23] Nurmaini S, Rachmatullah MN, Sapitri AI, Darmawahyuni A, Tutuko B, Firdaus F, et al. Deep learning-based computer-aided fetal echocardiography: Application to heart standard view segmentation for congenital heart defects detection. Sensors 2021;21.
[24] Wang Y, Huang W, Sun F, Xu T, Rong Y, Huang J. Deep multimodal fusion by channel exchanging. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H, editors. Advances in neural information processing systems,Vol. 33. Curran Associates, Inc.; 2020, p. 4835–45.
[25] Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Lecture notes in computer science (lecture notes in artificial intelligence and lecture notes in bioinformatics), vol. 11211, Springer Verlag; 2018, p. 833–51.
[26] Raza H, Ravanbakhsh M, Klein T, Nabi M. Weakly supervised one shot segmentation. In: 2019 IEEE/CVF international conference on computer vision workshop. 2019, p. 1401–6.
[27] Liu R, Lehman J, Molino P, Such FP, Frank E, Sergeev A, et al. Failing of convolutional neural networks and the CoordConv solution. Adv Inform Process Syst 2018;2018:9605–16.
[28] Sklansky J. Finding the convex hull of a simple polygon. Pattern Recognit Lett 1982;1(2):79–83.
[29] Wu S-T, Marquez M. A non-self-intersection Douglas-Peucker algorithm. In: 16th Brazilian symposium on computer graphics and image processing. 2003, p. 60–6.
[30] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst 2017;2017-December.
[31] Tongsong T, Tongprasert F, Srisupundit K, Luewan S. The complete three-vessel view in prenatal detection of congenital heart defects. Prenat Diagn 2010;30(1):23–9.
[32] Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: 3rd international conference on learning representations, ICLR 2015 - conference track proceedings. International Conference on Learning Representations, ICLR; 2014.