Hierarchical Spatio-temporal Segmentation Network for Ejection Fraction Estimation in Echocardiography Videos

Dongfang Wang, Jian Yang (✉), Yizhe Zhang, and Tao Zhou (✉)

PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
{dongfangwang,csjyang}@njust.edu.cn {yizhe.zhang.cs,taozhou.ai}@gmail.com

Abstract. Automated segmentation of the left ventricular endocardium in echocardiography videos is a key research area in cardiology. It aims to provide accurate assessment of cardiac structure and function through Ejection Fraction (EF) estimation. Although existing studies have achieved good segmentation performance, their results do not perform well in EF estimation. In this paper, we propose a Hierarchical Spatio-temporal Segmentation Network (HSS-Net) for echocardiography video, aiming to improve EF estimation accuracy by synergizing local detail modeling with global dynamic perception. The network employs a hierarchical design, with low-level stages using convolutional networks to process single-frame images and preserve details, while high-level stages utilize the Mamba architecture to capture spatio-temporal relationships. The hierarchical design balances single-frame and multi-frame processing, avoiding issues such as local error accumulation when relying solely on single frames or neglecting details when using only multi-frame data. To overcome local spatio-temporal limitations, we propose the Spatio-temporal Cross Scan (STCS) module, which integrates long-range context through skip scanning across frames and positions. This approach helps mitigate EF calculation biases caused by ultrasound image noise and other factors. We achieved state-of-the-art results on three datasets. Our code is available at https://github.com/DF-W/HSS-Net.

Keywords: Echocardiography · Ejection fraction · Segmentation

1 Introduction

The primary of echocardiography analysis is to assess cardiac function, where accurate segmentation of the left ventricular endocardium is essential to measure the heart's Ejection Fraction (EF) [3,1]. However, achieving automatic echocardiography segmentation presents significant challenges. Firstly, the quality of ultrasound images is often compromised by noise, artifacts, and boundary blurring [19,18]. Secondly, the heart undergoes complex motion and deformation during each heartbeat cycle.

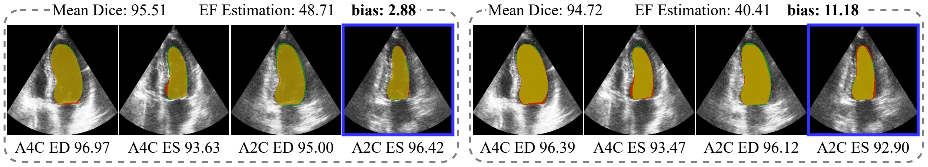


Fig. 1. Left ventricular segmentation maps for ED and ES frames from A2C and A4C views, with Dice segmentation metrics displayed at the bottom. **It is evident that smaller segmentation errors in the key regions (base and apex) within the blue box lead to greater EF calculation deviations**. The green, red, and yellow represent the ground truth, prediction, and overlapping regions, respectively.

Moreover, physicians typically only annotate the End-Diastolic (ED) and End-Systolic (ES) key frames, resulting in limited and sparse annotation data. Therefore, the automated segmentation method is trained on limited annotated data, providing a reliable foundation for the accurate evaluation of cardiac function clinical metrics.

Recently, deep learning methods for echocardiography video segmentation have emerged rapidly. Although effectiveness achieved, their results are not satisfactory when applied to ejection fraction estimation [16,22,21,24]. Many studies focus on 2D segmentation of ED and ES frames, with numerous unlabeled frames left unused, resulting in an inability to capture the continuity of cardiac motion [12,8,27]. To address this, Painchaud *et al.* [13] proposed a post-processing framework that leverages cardiac anatomical priors to enhance inter-frame consistency. However, its performance is highly dependent on the quality of the initial segmentation. Similarly, Wei *et al.* [20] used generated pseudo-labels for collaborative learning of segmentation and tracking, while the method is constrained by the quality of the pseudo-labels. The Transformer architecture, with multi-head self-attention [17], has been widely adopted in video object segmentation [2]. Temporal continuity across frames offers valuable segmentation cues. To exploit this, some methods add specialized modules atop self-attention to capture temporal information. However, this may cause over-reliance on multi-frame relations while overlooking fine details. Moreover, modeling long sequences with many labeled frames introduces heavy computational costs.

In this paper, we propose a Hierarchical Spatio-temporal Segmentation Network (HSS-Net) for echocardiography video segmentation, enhancing the precision of ejection fraction estimation. The low-level stages utilize convolutional networks to process single-frame images, preserving fine details, while the high-level stages employ the Mamba architecture to capture spatio-temporal relationships across multiple frames. By better integrating local and global features, the model reduces segmentation errors in key regions (*e.g.*, the base and apex) and mitigates volume calculation biases caused by local errors, as shown in Fig. 1. The network captures inter-frame motion patterns, enhancing the temporal consistency of the segmentation results and improving EF estimation stability. To fully utilize unlabeled data and strengthen the capture of dynamic cues, we propose a

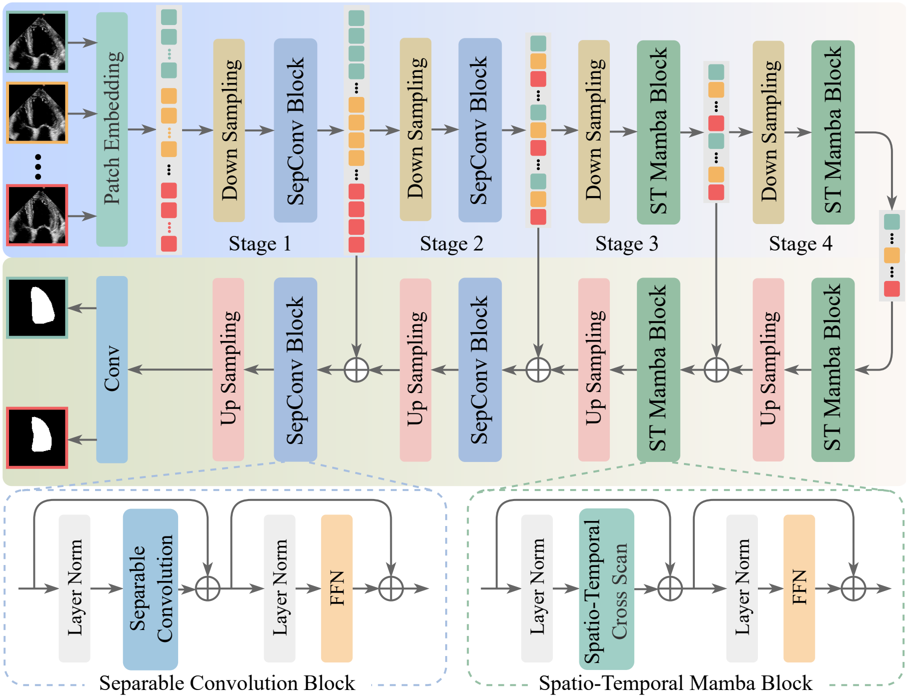


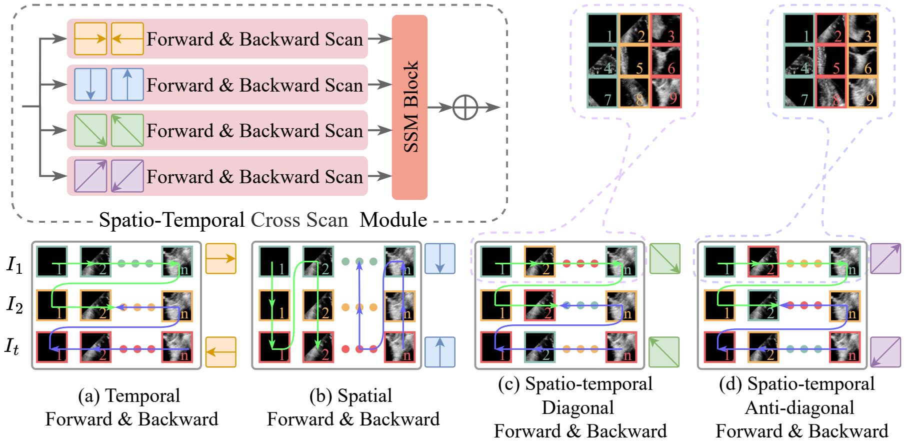
Fig. 2. Illustration of the proposed HSS-Net framework, which is a symmetric Encoder-Decoder architecture.

spatio-temporal cross scan module, which captures dynamic cues from different spatio-temporal perspectives. This module integrates long-range dependencies, such as apex motion changes and the correlation of lateral wall contraction, through a skip-connections-based spatio-temporal scan. Additionally, global dynamic modeling reduces sensitivity to interference factors, preventing abnormal jitter in segmentation boundaries and ensuring the reliability of EF estimation.

Contributions: i) We propose a hierarchical spatio-temporal segmentation framework, which combines convolutional and Mamba architectures for detailed local and temporal processing. **ii)** A STCS Module is proposed to capture dynamic cues from multiple perspectives, enhancing robustness and accuracy. **iii)** We present a novel jump scanning mechanism, which breaks local correlations to integrate global information, improving generalization across diverse samples. **iv)** Experimental results show the effectiveness of the proposed model.

2 Methodology

Overview. The architecture of our HSS-Net, illustrated in Fig. 2, consists of two modules: Encoder and Decoder. The first and second stages of the Encoder are primarily composed of stacked separable convolution blocks for low-level



(a) Temporal Forward & Backward

(b) Spatial Forward & Backward

(c) Spatio-temporal Diagonal Forward & Backward

(d) Spatio-temporal Anti-diagonal Forward & Backward

Fig. 3. Spatio-temporal cross scan module and its scanning sequence (I_t denotes the t-th frame in the video clip, and n denotes the number of patch sequences).

feature capture within single-frame images. The third and fourth stages are mainly composed of stacked spatio-temporal Mamba blocks for high-level feature capture across multiple frames. The Decoder architecture is symmetric to the Encoder, with a different number of stacked blocks. It is used to fuse multi-scale features and predict the segmentation mask. Specifically, given a video clip of T frames, denoted as $\mathbf{V} = \{I_1, I_2, ..., I_T\}$, we first apply patch embedding to divide these frames into different patches. The patch sequence is then fed into the encoder, resulting in the i-th stage feature $\mathbf{F_i}$, with a size of $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$, where H and W represent the height and width of the original frames, respectively, and $i \in \{1, 2, 3, 4\}$. Finally, these multi-scale features are passed to the decoder, where operations such as inter-frame perception, intra-frame perception, and up-sampling are performed to generate the predicted segmentation results.

2.1 Separable Convolution Block

Fig. 2 illustrates the structure of the separable convolution block. For the i-stage feature embedding of low-level $\mathbf{F_i} \in \mathbb{R}^{T\times C_i \times H_i \times W_i}, i \in \{1, 2\}$ of the given video clip, layer normalization $(\mathcal{LN}(\cdot))$ is performed before performing separable convolution operations $(\mathcal{SC}(\cdot))$. We follow the inverted separable convolution module from MobileNetV2 [15]. Subsequently, feeding the result into the FeedForward Network $(\mathcal{FFN}(\cdot))$ layer to capture low-level features such as edges and textures in single-frame images. The output features are either passed to the next stacked block or forwarded to the next stage. This process can be expressed as:

$$\mathbf{F_i} = \mathcal{SC}(\mathcal{LN}(\mathbf{F_i})) + \mathbf{F_i}, \quad \mathbf{F_i} = \mathcal{FFN}(\mathcal{LN}(\mathbf{F_i})) + \mathbf{F_i}. \tag{1}$$

2.2 Spatio-temporal Mamba Block

The structure of the spatial-temporal Mamba block is illustrated in Fig. 2. For the i-stage feature embedding of high-level $\mathbf{F_i} \in \mathbb{R}^{T \times C_i \times H_i \times W_i}, i \in \{3, 4\}$ of the given video clip, we transpose the channel and time dimensions and flatten the spatio-temporal feature embedding into a one-dimensional long sequence $\mathbf{s_i} \in \mathbb{R}^{C_i \times T H_i W_i}$. This sequence $\mathbf{s_i}$ is then fed into the spatio-temporal cross scan ($\mathcal{STCS}(\cdot)$) module and FFN layers. The STCS module establishes long-range dependencies both inter-frame and intra-frame from different spatio-temporal perspectives. This process can be defined as follows:

$$\mathbf{s_i} = \mathcal{STCS}(\mathcal{LN}(\mathbf{s_i})) + \mathbf{s_i}, \quad \mathbf{s_i} = \mathcal{FFN}(\mathcal{LN}(\mathbf{s_i})) + \mathbf{s_i}. \tag{2}$$

Finally, the output feature sequences are reshaped back to their original shape, and after down-sampling, the feature embeddings are passed to the next stage.

Spatio-temporal Cross Scan Module: As shown in Fig. 3, the STCS module with state space model (S6) [6] is designed for the spatio-temporal sequence modeling of video frames. It selectively scans the input sequence from various spatio-temporal perspectives, capturing intricate spatio-temporal relationships and providing a comprehensive understanding of the context.

To better understand and explore the spatio-temporal relationships among frames, we first unfold each frame's patches into sequences along rows and columns. As illustrated in Fig. 3, the patches of each frame are unfolded along rows to form temporal sequences, while the patches at the same position in different frames are unfolded along columns to form spatial sequences. The STCS module offers four different scanning modes: temporal, spatial, spatio-temporal diagonal, and spatio-temporal anti-diagonal. As depicted in Fig. 3(a), the module scans simultaneously along the temporal sequence in both forward and backward directions to explore bidirectional temporal dependencies. In Fig. 3(b), the module scans simultaneously along the spatial sequence in both upward and downward directions to explore bidirectional spatial dependencies. The selective spatio-temporal scanning explicitly considers both intra-frame and inter-frame coherencies and leverages the SSM to establish long-range dependencies of intra-frame and inter-frame.

The heart motion during a heartbeat is not fully synchronous, leading to similarities and differences in information among video frames. To disrupt local data correlations, integrate global information, capture diverse features, and enhance the model's generalization and understanding of cardiac structure and motion, we propose a new spatio-temporal diagonal and anti-diagonal scanning method (see Fig. 3(c) and (d)). We rearrange spatial sequence positions into diagonal and anti-diagonal patterns and scan temporally in forward and backward directions. This cross-frame and cross-position scanning improves the model's ability to integrate global information and understand cardiac motion and structure.

Loss Function: During training, our loss function includes the Dice loss $\mathcal{L}_{dice}$ [11] and binary cross-entropy loss $\mathcal{L}_{bce}$. Thus, the total loss function $\mathcal{L}_{total} = \alpha\mathcal{L}_{dice}(P, G) + (1 - \alpha)\mathcal{L}_{bce}(P, G)$, where G denotes the ground-truth, P denotes the predicted masks, and the balance weight $\alpha = 0.8$ in our experiments.

Table 1. The quantitative results on the CAMUS. FLOPs represent the average computational complexity per frame at a size of 256 × 256.

Methods	Params	FLOPs	corr	bias± std	Dice	HD95
UNet++ [28]	26.9M	37.7G	81.68	6.05±6.81	91.87	16.16
TransUNet [4]	105.3M	38.6G	86.22	**1.72±6.07**	92.73	13.71
SegFormer [23]	47.4M	20.9G	79.10	7.84±7.21	91.17	18.46
H2Former [7]	33.7M	33.1G	82.35	5.79±6.87	91.78	16.06
SSCF [22]	53.7M	15.1G	84.48	4.11±6.37	92.59	14.18
PKEchoNet [21]	25.7M	7.2G	76.20	4.13±8.45	93.02	12.93
VideoMamba [10]	75.6M	22.0G	75.21	8.02±8.00	91.53	16.43
Vivim [25]	59.6M	20.6G	78.09	5.75±7.35	92.79	12.74
HSS-Net (Ours)	31.2M	5.6G	**90.47**	2.43±5.02	**93.89**	**11.29**

Table 2. The quantitative results on the EchoNet-Pediatric and EchoNet-Dynamic datasets. The HD95 metrics are reported in pixels.

Methods	EchoNet-Pediatric				EchoNet-Dynamic			
	corr	bias± std	Dice	HD95	corr	bias± std	Dice	HD95
UNet++ [28]	69.33	7.62±10.16	90.73	3.65	73.91	9.43±9.02	91.50	2.99
TransUNet [4]	73.09	6.54±9.81	91.11	3.52	74.17	4.67±9.51	91.92	2.96
SegFormer [23]	66.72	6.25±10.77	91.10	3.52	73.12	7.07±9.37	92.07	2.90
H2Former [7]	69.77	6.24±10.06	90.89	3.58	74.78	6.10±9.23	91.68	3.12
SSCF [22]	63.29	5.28±11.82	91.07	3.50	74.87	6.26±9.16	92.35	2.80
PKEchoNet [21]	65.04	6.23±11.21	91.00	3.57	75.43	4.34±9.50	92.45	2.71
VideoMamba [10]	67.34	6.39±11.39	91.06	3.50	78.62	4.50±8.29	92.48	2.75
Vivim [25]	69.92	5.59±10.31	91.12	3.46	81.12	7.02±7.47	92.46	2.73
HSS-Net (Ours)	**76.91**	**1.29±8.68**	**91.90**	**3.23**	**84.50**	**0.95±6.75**	**92.67**	**2.66**

3 Experiments

Datasets: In this study, three publicly available echocardiography video datasets are used, namely CAMUS [9], EchoNet-Pediatric [14], and EchoNet-Dynamic [24]. • **CAMUS** comprises 500 cases acquired at the University Hospital of St Etienne (France), each including 2D apical 2-chamber and 4-chamber view videos, with annotations provided for all frames. • **EchoNet-Pediatric** is collected from Lucile Packard Children's Hospital Stanford, including 7,643 video clips from 1,958 patients aged 0 to 18 years. This dataset consists of either parasternal short axis or apical 4-chamber views, with only the ED and ES frames annotated. • **EchoNet-Dynamic** consists of 10,030 apical 4-chamber view echocardiography videos collected from Stanford University Hospital, with only ED and ES frames annotated for each video. We uniformly sample 10 frames of each video clip from datasets, following previous research [5,21]. The video clips are cropped to ensure that the ED frame is the first and the ES frame is the last, thereby capturing a complete heartbeat cycle. The frame size was adjusted to 256 × 256,

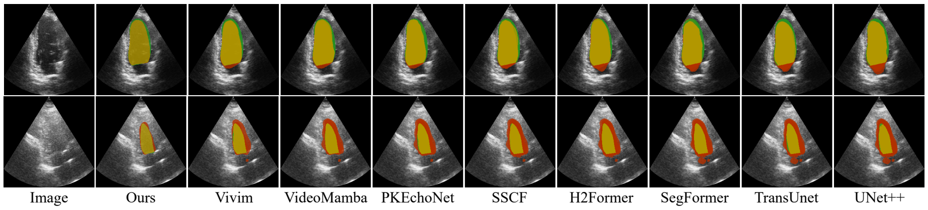


Fig. 4. Visualization of segmentation maps for different models.

and only the annotations from the ED and ES frames were used for training and evaluation. For the EchoNet-Dynamic dataset, we used the original dataset splits. For the other datasets, we follow related studies [24,26] and split the data into training, validation, and testing sets in an 8:1:1 ratio.

Evaluation Metrics: We report three statistical metrics for left ventricular ejection fraction. The estimation methods vary due to different views provided by the datasets. For the EchoNet-Dynamic and EchoNet-Pediatric datasets, both ground truth and predicted ejection fractions are obtained using the Simpson's single-plane method of disks [9]. For the CAMUS dataset, the Simpson's biplane method of disks [24] is used to calculate ejection fractions. We follow [5,21] and calculate the Pearson correlation coefficient (corr), mean bias (bias), and standard deviation (std) for the predicted and ground truth ejection fractions. Additionally, we employed two widely used segmentation evaluation metrics: the mean Dice coefficient (Dice) and Hausdorff Distance at 95% (HD95).

Implementation Details: Our model is implemented using PyTorch and is trained or inferred on two NVIDIA RTX 4090 GPU. We train our model end-to-end using the Adam optimizer and employ the cosine annealing strategy to adjust the learning rate. The maximum and minimum learning rates are set to 1e-4 and 1e-5, respectively, and the maximum training epoch is set to 120. During training, we apply gamma augmentation, random scaling, random rotation, and random contrast adjustments, each with a probability of 0.5.

3.1 Comparison with State-of-the-art Methods

Quantitative Comparisons: The quantitative results for the three datasets are presented in Tables 1 and 2. It can be seen that our model outperforms other methods in all datasets. However, image-based methods (The first four) still exhibit competitive performance on certain datasets. These methods focus on capturing local features within single-frame images, whereas video-based methods (The last four), although proficient in capturing temporal information, may not fully exploit their advantages on datasets with minimal inter-frame differences or subtle dynamic changes. Our model hierarchically processes single-frame and multi-frame information, balancing both detailed features and dynamic changes in echocardiography. This approach is particularly advantageous for ejection fraction estimation. Additionally, it maintains an optimal balance between model performance and computational complexity.

Table 3. Quantitative results of ablation studies.

Settings	CAMUS				EchoNet-Dynamic			
	corr	bias± std	Dice	HD95	corr	bias± std	Dice	HD95
Image-level	83.48	5.28±6.71	92.93	13.91	74.79	6.00±9.34	91.93	2.97
Video-level	80.67	4.48±7.04	93.04	13.09	78.01	4.62±7.92	92.03	2.91
w/o Temporal	83.73	4.20±6.67	93.02	14.26	78.42	4.80±8.02	92.10	2.90
w/o Spatio	80.44	5.05±7.44	93.26	12.57	77.86	4.76±8.21	92.32	2.79
w/o ST Diagonal	86.69	4.05±5.99	93.21	12.40	79.65	6.03±7.70	92.12	2.84
w/o ST Anti-diagonal	88.09	2.95±5.57	93.27	12.28	81.11	4.85±7.54	92.26	2.82
HSS-Net (Ours)	**90.47**	**2.43±5.02**	**93.89**	**11.29**	**84.50**	**0.95±6.75**	**92.67**	**2.66**

Qualitative Comparisons: We present visualizations of several challenging cases. As shown in Fig. 4, these sample images exhibit artifacts, speckle noise, and blurred boundaries. Such challenging conditions mislead most of the compared models, resulting in missed or misclassified regions. In contrast, our model accurately locates the regions and delineates the boundaries. These visualizations further demonstrate that our approach can achieve better segmentation results and robustly handle poor-quality images.

3.2 Ablation Study

Effectiveness of Hierarchical Design: To validate the effectiveness of the hierarchical design, we perform two sets of ablation experiments. One set processes single-frame images at all stages using only separable convolution blocks (labeled as Image-level). The other set processes multi-frame images at all stages using only spatio-temporal Mamba blocks (labeled as Video-level). As shown in Table 3, the model performance in both experiments decreased to varying degrees, especially in the key metric of Pearson correlation for ejection fraction estimation. This demonstrates that by extracting fine-grained details at the low level and modeling cross-frame temporal relationships at the high level, the model effectively handles subtle differences across different conditions, providing a more reliable foundation for clinical assessment of cardiac function.

Effectiveness of Spatio-temporal Cross Scan Module: We investigate the effect of each mode in the STCS module. As shown in Table 3, all performance metrics exhibit varying degrees of decline compared to our full method. Temporal scanning captures sequential dynamic information during cardiac motion, which is crucial to accurately modeling heart movement patterns. Spatial scanning aids the model in understanding changes at the same location across different time frames. This enhances the model's ability to perceive spatial consistency features. Spatio-temporal diagonal and anti-diagonal scanning effectively capture complex interactions between temporal and spatial dimensions, enhancing the model's ability to integrate spatio-temporal information.

4 Conclusion

We propose a novel method HSS-Net that employs a hierarchical design. The low-level stages use convolutions to extract local details from single-frame images, while the high-level stages leverage the Mamba architecture to process spatio-temporal information across multiple frames. By handling single-frame and multi-frame information hierarchically, the model's accuracy and robustness are enhanced. Extensive experimental results demonstrate that our method achieves state-of-the-art results on three benchmark datasets.

Acknowledgement. This work was in part supported by the National Science Fund of China under Grant Nos. 62172228, U24A20330, and 62361166670.

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

References

1. Akkus, Z., Aly, Y.H., Attia, I.Z., Lopez-Jimenez, F., Arruda-Olson, A.M., Pellikka, P.A., Pislaru, S.V., Kane, G.C., Friedman, P.A., Oh, J.K.: Artificial intelligence (ai)-empowered echocardiography interpretation: a state-of-the-art review. Journal of Clinical Medicine **10**(7), 1391 (2021)
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: International Conference on Computer Vision. pp. 6836–6846 (2021)
3. Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., Rueckert, D.: Deep learning for cardiac image segmentation: a review. Frontiers in Cardiovascular Medicine **7**, 25 (2020)
4. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
5. Deng, X., Wu, H., Zeng, R., Qin, J.: Memsam: Taming segment anything model for echocardiography video segmentation. In: Computer Vision and Pattern Recognition. pp. 9622–9631 (2024)
6. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
7. He, A., Wang, K., Li, T., Du, C., Xia, S., Fu, H.: H2former: An efficient hierarchical hybrid transformer for medical image segmentation. IEEE Transactions on Medical Imaging **42**(9), 2763–2775 (2023)
8. Leclerc, S., Smistad, E., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Belhamissi, M., Israilov, S., Grenier, T., et al.: Lu-net: a multi-stage attention network to improve the robustness of segmentation of left ventricular structures in 2-d echocardiography. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control **67**(12), 2519–2530 (2020)
9. Leclerc, S., Smistad, E., Pedrosa, J., Ostvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., Lartizien, C., Dhooge, J., Lovstakken, L., Bernard, O.: Deep learning for segmentation using an open large-scale

dataset in 2d echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (2019)

10. Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., Qiao, Y.: Videomamba: State space model for efficient video understanding. arXiv preprint arXiv:2403.06977 (2024)
11. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Fourth International Conference on 3D Vision. pp. 565–571 (2016)
12. Moradi, S., Oghli, M.G., Alizadehasl, A., Shiri, I., Oveisi, N., Oveisi, M., Maleki, M., Dhooge, J.: Mfp-unet: A novel deep learning based approach for left ventricle segmentation in echocardiography. Physica Medica **67**, 58–69 (2019)
13. Painchaud, N., Duchateau, N., Bernard, O., Jodoin, P.M.: Echocardiography segmentation with enforced temporal consistency. IEEE Transactions on Medical Imaging **41**(10), 2867–2878 (2022)
14. Reddy, C.D., Lopez, L., Ouyang, D., Zou, J.Y., He, B.: Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients. Journal of the American Society of Echocardiography **36**(5), 482–489 (2023)
15. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Computer Vision and Pattern Recognition. pp. 4510–4520 (2018)
16. Thomas, S., Gilbert, A., Ben-Yosef, G.: Light-weight spatio-temporal graphs for segmentation and ejection fraction prediction in cardiac ultrasound. In: Medical Image Computing and Computer Assisted Intervention. vol. 13434, pp. 380–390 (2022)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. vol. 30 (2017)
18. Wang, D., Zhou, T., Yang, J.: Hybrid-frequency feature evolution network for endoscopic ultrasound image segmentation. In: International Symposium on Biomedical Imaging. pp. 1–5 (2025)
19. Wang, D., Zhou, T., Zhang, Y., Gao, S., Yang, J.: Frequency-aware interaction network for ultrasound image segmentation. IEEE Transactions on Circuits and Systems for Video Technology pp. 1–1 (2024)
20. Wei, H., Ma, J., Zhou, Y., Xue, W., Ni, D.: Co-learning of appearance and shape for precise ejection fraction estimation from echocardiographic sequences. Medical Image Analysis **84**, 102686 (2023)
21. Wu, H., Lin, J., Xie, W., Qin, J.: Super-efficient echocardiography video segmentation via proxy-and kernel-based semi-supervised learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2803–2811 (2023)
22. Wu, H., Liu, J., Xiao, F., Wen, Z., Cheng, L., Qin, J.: Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion. Medical Image Analysis **78**, 102397 (2022)
23. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. vol. 34, pp. 12077–12090 (2021)
24. Yang, J., Ding, X., Zheng, Z., Xu, X., Li, X.: Graphecho: Graph-driven unsupervised domain adaptation for echocardiogram video segmentation. In: International Conference on Computer Vision. pp. 11844–11853 (2023)
25. Yang, Y., Xing, Z., Zhu, L.: Vivim: a video vision mamba for medical video object segmentation. arXiv preprint arXiv:2401.14168 (2024)

26. Ye, Z., Chen, T.: P-mamba: Marrying perona malik diffusion with mamba for efficient pediatric echocardiographic left ventricular segmentation. arXiv preprint arXiv:2402.08506 (2024)
27. Zamzmi, G., Rajaraman, S., Hsu, L.Y., Sachdev, V., Antani, S.: Real-time echocardiography image analysis and quantification of cardiac indices. Medical Image Analysis **80**, 102438 (2022)
28. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Transactions on Medical Imaging **39**(6), 1856–1867 (2019)