IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 May 2024, accepted 4 July 2024, date of publication 8 July 2024, date of current version 11 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3424546

RESEARCH ARTICLE

HFE-Mamba: High-Frequency Enhanced Mamba Network for Efficient Segmentation of Left Ventricle in Pediatric Echocardiograms

ZI YE[1,2], TIANXIANG CHEN[3], DAN WANG[4], FANGYIJIE WANG[5], AND LIJUN ZHANG[6]

[1]Institute of Intelligent Software, Guangzhou 325035, China
[2]Automotive Software Innovation Center, Chongqing 100190, China
[3]School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230026, China
[4]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China
[5]School of Medicine, University College Dublin, Dublin 4, D04 V1W8 Ireland
[6]Institute of Software, Chinese Academy of Sciences, Beijing 100045, China

Corresponding author: Lijun Zhang (zhanglj@ios.ac.cn)

ABSTRACT Automated ventricular function analysis can enhance healthcare consistency and accessibility, particularly in resource-limited settings. Current segmentation methods trained on adult heart ultrasounds struggle to accurately outline the irregular shape of the left ventricle owing to their limited exploration of border features. HFE-Mamba is introduced for left ventricle segmentation with shape awareness in order to address the existing challenge. Therefore, this proposal introduces the High-Frequency Enhancement Block (HFEB), enhancing the high-frequency component of left ventricles, particularly the boundary area in pediatric echocardiograms. Moreover, this also facilitates the investigation of target boundary specifics while extracting features. The incorporation of newly suggested vision mamba layers into encoder and decoder branches enhances the model's computational and memory efficiency while capturing global dependencies. Tests conducted on two publicly available datasets indicate the superior predictive accuracy of the HFE-Mamba model in identifying target shapes.

INDEX TERMS Mamba, high frequency, left ventricle, echocardiogram, segmentation.

I. INTRODUCTION

Congenital heart diseases (CHD), known for their significant impact on mortality and morbidity, has long been a critical health concern [1]. Due to its portability, affordability, and real-time capabilities, the Echocardiogram is essential in clinics to detect and treat CHD in children. Besides, accurate segmentation of cardiac structure in echocardiography images is a critical step for different analytical and diagnostic processes [2]. A schematic illustration of manually delineating anatomical structures in an echocardiogram is displayed in **Figure 1**.

Among these indices, the Left Ventricular Ejection Fraction (LVEF) is the most commonly used and vital metric for assessing systolic function [3]. The LVEF is predominantly calculated using the biplane Simpson's standard protocol method in clinical settings [4]. This technique involves the manual delineation of the left ventricular endocardium by physicians in specific frames of the apical two-chamber (A2C) and apical four-chamber (A4C) echocardiographic views, a process essential for the identification of the end-systolic volume (LVESV) and end-diastolic volume (LVEDV). However, echocardiography often encounters challenges including significant speckle noise, limiting its imaging technique. This issue can lead to blurred boundaries and potential artifacts in the cardiac tissue imagery, making the tracing process more complex and dependent on the physician's expertise. Therefore, this manual approach to calculating LVEF is prone to errors, which can significantly affect the reliability of the diagnostic outcome, highlighting the need for precision and care in this critical measurement.

Machine learning and artificial intelligence have significantly improved the dependability and precision of evaluating left ventricular (LV) function using echocardiography in

The associate editor coordinating the review of this manuscript and approving it for publication was Nuno M. Garcia.

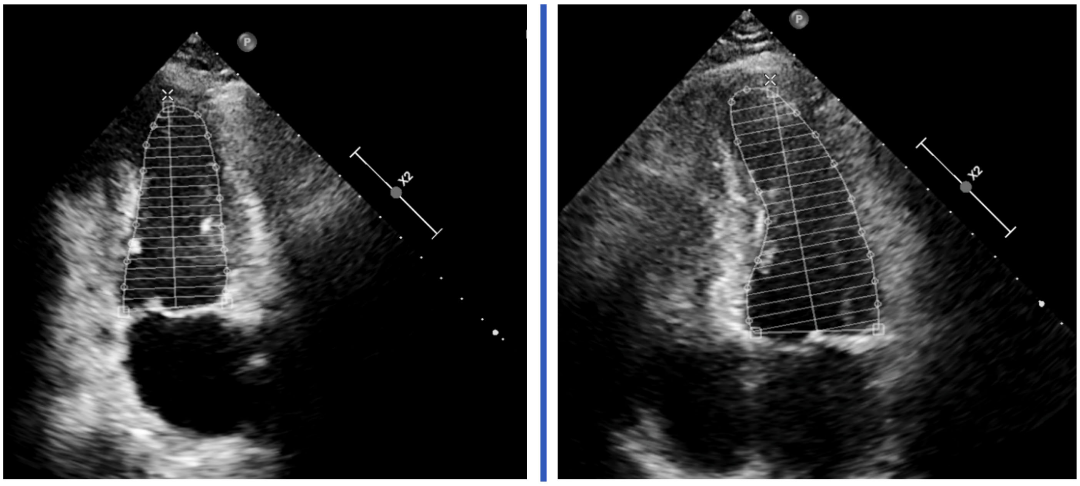


FIGURE 1. **The examples of manually delineating anatomical structures in an echocardiogram.**

adults, as shown by multiple research projects. Machine learning is more difficult in youngsters due to diverse anatomical anomalies, heart rate, stature, and cooperative capacity. Various factors influence the spatial and temporal resolutions, eventually influencing echocardiographic imaging quality [5]. As a result, there is concern about how well machine learning models built on adult datasets can be applied in pediatric echocardiography owing to the more significant variabilities.

Numerous transformer topologies have been investigated following the Vision Transformer's (ViT) success in medical image tasks [6]. Swin-Unet [7] refers to a transformer-based U-shaped Encoder-Decoder network that has been one of the most vital techniques. Nevertheless, the self-attention mechanism in Transformers presents difficulties in speed and memory consumption when dealing with distant visual relationships, including exploring high-resolution images. Recently, state space models (SSMs) have exhibited significant potential for long-sequence modeling. Mamba [8] is a modern state space model excelling in capturing long-range relationships compared to transformers. This hardware-aware model is designed with linear complexity to guarantee efficient training and inference operations.

The significance of frequency domain analysis in computer vision is well-documented in the literature [9], [10], emphasizing that low frequencies in images represent global structures and color while high frequencies expose particular features. We draw inspiration from frequency learning research to develop HFE-Mamba, highlighting high-frequency boundary details. Our HFE-Mamba includes a unique module, the High-Frequency Enhancement Block (HFEB), that concentrates on high-frequency (HF) data, which is essential for identifying object boundaries in segmentation assignments. Our contributions can be outlined as follows:

- High-Frequency Enhancement Blocks (HFEB) are incorporated into our network encoder to emphasize the high-frequency elements of the retrieved feature maps, which correspond to the boundaries of segmented objects, enhancing shape-aware segmentation accuracy during feature extraction.
- The HFEB and local feature extraction are integrated with Mamba's global dependency modeling to obtain high performance on two distinct pediatric echocardiogram datasets.
- Experiments demonstrate that our HFE-Mamba performs exceptionally well on two public datasets compared to other recently successful segmentation algorithms.

II. RELATED WORK

Till the present, deep learning (DL) development has promoted automatic medical image segmentation, and several well-known deep learning frameworks have provided good ideas for echocardiography segmentation with outstanding performance.

A. CONVENTIONAL LEFT VENTRICLE SEGMENTATION

Recently, innovative deep-learning methods have become increasingly vital in medical analysis and represent the forefront of leveraging complex neural networks for precise left ventricular segmentation in echocardiography, as summarized in **Table 1**. MAEF-Net [11] employs a multi-attention mechanism and spatial pyramid feature fusion, significantly improving the accuracy. Similarly, SegCaps [12] introduces an optimized capsule-based network, addressing conventional CNNs' limitations through effectively capturing spatial information and part-whole relationships with fewer parameters. EchoEFNet [14], another multi-task network, leverages ResNet50 with dilated convolution. It segments the left ventricle and identifies landmarks, which can therefore facilitate the automatic calculation of the left ventricular ejection fraction (LVEF) using the biplane Simpson's method.

TABLE 1. Recent innovative deep-learning methods for precise left ventricular segmentation in echocardiography.

Model Name	Core Concept	Dataset	DSC
MAEF-Net [11]	Multi-Attention Mechanism	EchoNet-Dynamic	93.10%
SegCaps [12]	Conv-Deconv Capsule Model	CAMUS	84.48%
DL-based tool [13]	YOLOv7 and U-Net	CAMUS	92.63%
EchoEFNet [14]	Multi-task	CMUEcho	96.5%
		CAMUS	93.4%
Contrastive-echo [15]	SimCLR and BYOL	EchoNet-Dynamic	92.52%
		CAMUS	93.11%
Temporal Consistency [16]	2D+time CNNs	CAMUS	95.5%

A few researchers have been attempting to overcome data limitations, processing, and consistency challenges. One method combines YOLOv7 and U-Net for segmenting multiple anatomical structures in echocardiographic images [13]. Another self-supervised contrastive learning method was proposed [15], particularly effective when annotated images are scarce and boosts the performance of UNet and DeepLabV3 with minimal training data. Furthermore, a new framework, which addressed temporal inconsistency in CNN-based segmentation [16], has enhanced 2D+time apical long-axis heart shape segmentation in echocardiography using a constrained autoencoder for spatiotemporal correction based on physiologically realistic heart shapes.

B. LEFT VENTRICLE SEGMENTATION IN PEDIATRIC CARDIOLOGY

AI modalities have also expanded feet to the specialty of pediatric cardiology. It was found that using neural networks and machine learning has significantly improved the LV segmentation of echocardiograms, which can thus augment the clinicians' diagnostic accuracy of pediatric heart diseases.

A novel algorithm presents a Dual Attention Enhancement Feature Fusion Network, integrating a dual-path feature extraction module (DP-FEM) for rich feature extraction and a high and low-level feature fusion module (HL-FFM) for semantic and spatial information, with a unique hybrid loss function to address pixel-level misalignment and boundary ambiguities [17]. Similarly, an Attention-Guided Dual-Path Network named AIDAN was introduced to deal with challenges including low signal-to-noise ratio and internal variability in heart appearance [18]. It employs a Convolutional Block Attention Module (CBAM) for discriminative feature learning and efficient spatial and contextual path fusion. Furthermore, the Multi-Scale Wavelet Network (MS-Net) algorithm focuses on hierarchical feature-guided fusion and uses Discrete Wavelet Transform to lower the impact of image noise [19]. The MS-Net combines bidirectional feature fusion (BFF-Net) and a wavelet-Unet (W-Unet) module, effectively integrating context and detail information for accurate segmentation.

While exhibiting novel algorithms for pediatric echocardiography segmentation, the studies above are limited by their reliance on proprietary datasets. The EchoNet-Peds dataset, developed by Stanford University, indicates the first publicly available pediatric echocardiography dataset [20], featuring 4,467 echocardiograms from 1,958 patients, including a 43% female demographic and ages ranging from newborns to 18 years. This comprehensive dataset yielded 7,643 video clips and 17,600 labeled images, primarily from A4C and PSAX view clips. As a result, the video clips were strategically allocated, with 6,114 (80%) for training, 765 (10%) for testing, as well as 764 (10%) for validation.

C. FREQUENCY DOMAIN ANALYSIS IN VISION

The frequency domain analysis has indicated a significant advancement in CV tasks. LITv2 [9], utilizing the HiLo mechanism, differentiates high and low-frequency details. This therefore enhances local and global features. Similarly, the Camouflaged Object Detection (COD) approach incorporates frequency domain clues, with key components Frequency Enhancement Module (FEM) and High-Order Relation Module (HOR) for better detection of objects camouflaged in their environments [10].

Innovations including PIDNet [21], SpectFormer [22], and SVT [23] further demonstrate this trend. PIDNet's three-branch design, inspired by Proportional-Integral-Derivative (PID) controller, adeptly handles detailed, contextual, and boundary information, which can ensure effective feature fusion. The SpectFormer merges spectral layers with attention mechanisms, showing adaptability across various tasks and datasets. With its spectrally scattering network, SVT adeptly manages attention complexity and captures fine-grained details, standing out in both efficiency and performance. These models are pivotal in the evolution of vision transformers, emphasizing the significance of integrating novel concepts for enhanced visual processing.

D. MAMBA

Mamba [8] is a novel deep sequence model architecture addressing the computational inefficiency of traditional

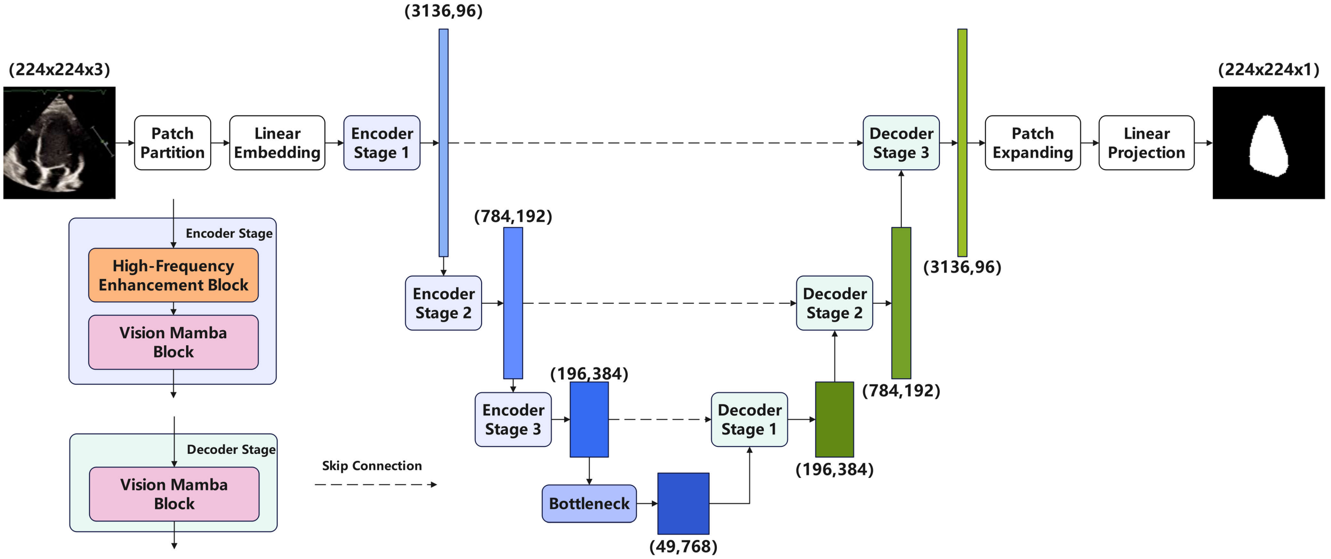


FIGURE 2. The overall structure of the proposed model HFE-Mamba.

Transformers on long sequences. It is based on selective state space models (SSMs), improving upon previous SSMs by allowing the model parameters to be input functions. The mathematical function of SSMs can be represented as follows:

$$x'(t) = Ax(t) + Bu(t) \tag{1}$$

$$y(t) = Cx(t) \tag{2}$$

in which, state matrix $A \in R^{N\times N}$ and $B, C \in R^N$ are its parameters, and $x(t) \in R^N$ denotes the implicit latent state.

This enables Mamba to selectively propagate or forget information alongside the sequence length dimension, depending on the current token. Despite this change preventing efficient convolutions, the architecture incorporates a hardware-aware parallel algorithm that operates in recurrent mode, keeping linear scaling in sequence length while achieving faster inference speeds.

Specifically, this hardware-aware algorithm utilizes a scan operation instead of convolution, effectively removing the requirement to create the extended state. To enhance GPU utilization and efficiently materialize the state h within the memory hierarchy, hardware-aware state expansion is enabled by selective scan. This design decision is vital because it eliminates unnecessary input/output access throughout the many layers of the GPU memory hierarchy, thus improving memory management and utilization.

Moreover, empirical evidence demonstrates that our technique achieves a speed improvement of up to thrice on contemporary GPUs compared to earlier methodologies. The substantial acceleration highlights the benefits of developing algorithms with a profound comprehension of the fundamental hardware architecture.

III. METHODOLOGY

As shown in **Figure 2**, the proposed HFE-Mamba design comprises an encoder, bottleneck, decoder, and skip connections. Skip connections combine the multi-scale features from the encoder with the up-sampled features, similar to the U-Net architecture. Moreover, we will provide a more detailed explanation of our model in the upcoming subsections.

A. HIGH-FREQUENCY ENHANCEMENT BLOCK

We provide a new approach called High-Frequency Enhancement Block (HFEB) to enhance the processing of high-frequency information for semantic segmentation tasks. **Figure 3** displays the structure of our approach.

The HFEB includes a Dual-Tree Complex Wavelet Transform (DTCWT) handling Low-Frequency (LF) and High-Frequency (HF) components. Then, these components are passed through the Tensor Blending Method (TBM) and Einstein Blending Method (EBM) to produce Low-Frequency Representation (LFR) and High-Frequency Representation (HFR), which are adjusted by trainable weight matrices W_ϕ and W_ψ [23].

This module employs the Dual-Tree Complex Wavelet Transform Inverse (DTCWT Inverse) to process only the High-Frequency Representation (HFR) through eliminating the low-frequency component, enhancing high-frequency details. An ‘Adding’ process combines outputs from the Cross-Attention and Self-Attention modules in order to emphasize fine-grained details and long-range dependencies in the data.

The Self-Attention mechanism isolates and processes the high-frequency component. With the application of this technique, the model can concentrate on particular regions of the image with notable textural details, therefore enhancing the importance of this information. The formulation of

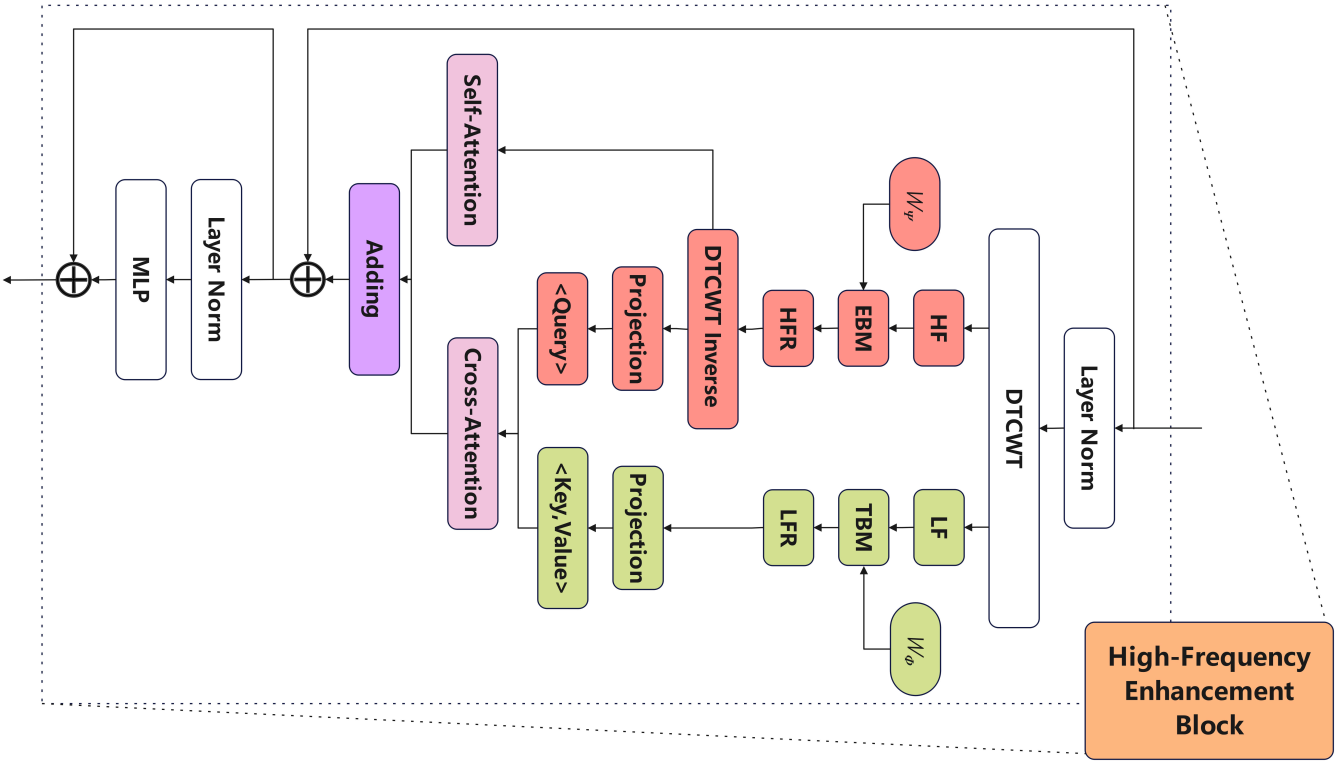


FIGURE 3. The structure of the high-frequency enhancement block.

Self-Attention is as follows:

$$\text{Self-Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

where Q, K, and V are the query, key, and value matrices derived from the HFR after the DTCWT Inverse procedure.

In the HFEB framework, the Cross-Attention mechanism combines signals from different frequency spectra. Queries derived from the high-frequency component are applied to modulate keys and values collected from the low-frequency domain, enabling the integration of global and fine-grained information. The simplified expression of Cross-Attention is:

$$\text{Cross-Attention}(Q_{HF}, K_{LF}, V_{LF}) = \text{softmax}\left(\frac{Q_{HF}K_{LF}^T}{\sqrt{d_k}}\right)V_{LF} \quad (4)$$

where Q_{HF} is the query matrix derived from the HFR, while K_{LF} and V_{LF} are the key and value matrices derived from the LFR, respectively.

B. VISION MAMBA BLOCK

A block of SSM handles the input tokens after passing via the HFEB. In this study, we drew inspiration from Vision Mamba [24], including bidirectional sequence modeling specifically tailored for vision applications. These designs incorporate modest yet crucial changes to the typical SSM block design.

Figure 4 displays the Vision Mamba Block (VMB) integrating forward and backward directions. It benefits from bidirectional state space modeling, providing data-dependent global visual context and comparable modeling capabilities for Transformers with reduced computational complexity.

C. ENCODING AND DECODING PATH

The encoder layer combines HFEB and VMB to collect high-frequency details and focus on relevant spatial regions. Blocks are stacked systematically to embed the input image into a latent space and conduct representation learning over consecutive stages.

The VMB plays a vital role in reconstructing spatial details from encoded characteristics within the decoder architecture. The skip connections, like in the Swin-Unet model [7], combine the multi-scale features from the encoder with the up-sampled characteristics. Shallow and deep features are combined to mitigate the loss of spatial information due to down-sampling.

IV. MATERIALS AND EVALUATION METRICS

The current section will begin by providing a thorough overview of the dataset utilized to assess the proposed model's effectiveness. Then, we will cover the implementation specifics and describe the assessment metrics used in our study.

A. DATASET

The dataset comprises echocardiographic evaluations from patients at Lucile Packard Children's Hospital Stanford

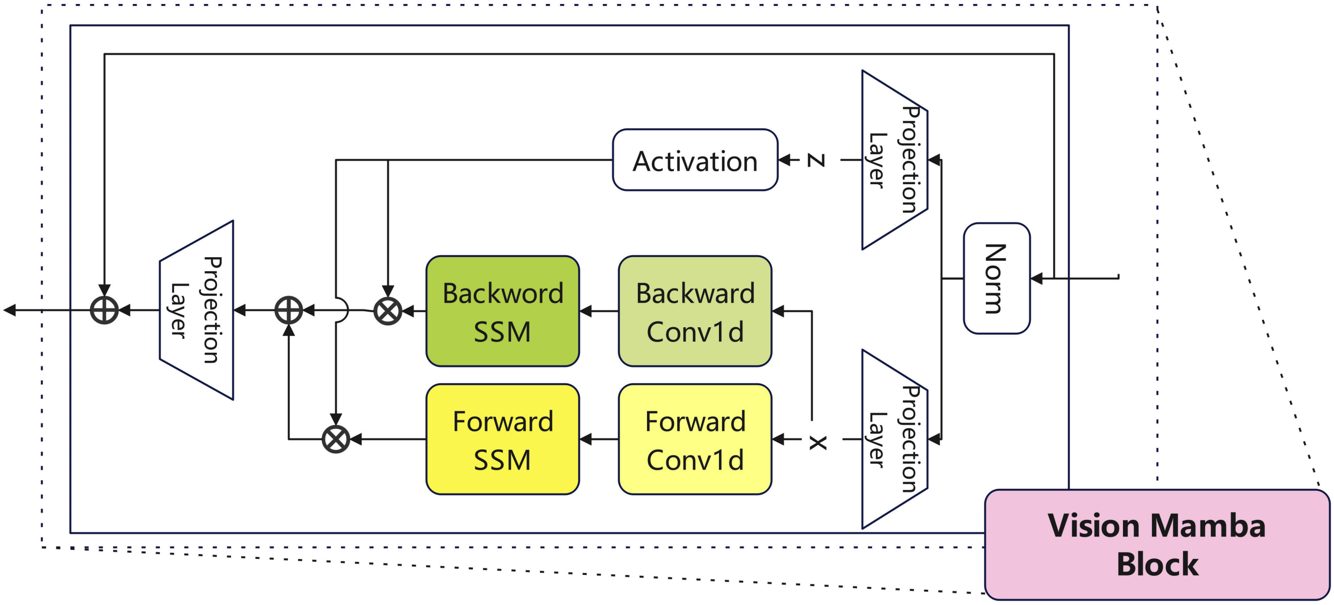


FIGURE 4. The structure of the vision mamba block.

TABLE 2. Performance comparison with other methods on both PSAX and A4C.

Methods	Dataset PSAX			Dataset A4C		
	Precision	Recall	DSC	Precision	Recall	DSC
UNet (FCN) [27]	0.8492	0.8761	0.8624	0.8279	0.8345	0.8312
UNet (PSPNet) [27]	0.8651	0.87	0.8675	0.8204	0.8607	0.8401
UNet (deeplabv3) [27]	0.8517	0.8854	0.8682	0.8076	0.8612	0.8336
ResNet-50 [28]	0.89	0.909	0.8994	0.8837	0.8773	0.8805
Swin-Unet [7]	0.8812	0.9010	0.9113	0.8722	0.8835	0.8920
Spectformer [22]	0.8932	0.9063	0.9160	0.8909	0.9035	0.9009
PVT [29]	0.867	0.8366	0.8515	0.7166	0.6553	0.6846
Uniformer [30]	0.8963	0.9016	0.8910	0.8824	0.8885	0.8820
EfficientVit [31]	0.9100	0.9134	0.9073	0.8969	0.8915	0.8918
Flatten Transformer [32]	0.9215	0.9122	0.9168	0.913	0.8832	0.8978
MaxVit [33]	0.9024	0.9249	0.9135	0.8912	0.9066	0.8988
Ours	0.9200	**0.9290**	**0.9188**	**0.9149**	**0.9103**	**0.9029**

from 2014 to 2021, authorized by the Stanford University Institutional Review Board. The dataset contains totally 4,467 echocardiograms collected from 1,958 patients, 43% female, aged between 0 and 18 years (mean ± SD: 10 ± 5.4 years). The patients were classified into two groups based on their echocardiographic results: those with structurally normal hearts and average ejection fraction (EF) and those with structurally normal hearts but systolic dysfunction (including dilated cardiomyopathy, chemotherapy-induced systolic dysfunction) without congenital heart disease [20].

Echocardiograms were performed with Philips iE33, Siemens Acuson SC2000, or Philips Epiq 7 ultrasound machines. A Siemens Syngo Dynamics picture archiving and communication system were used to save and examine the videos. After additional processing, the dataset was employed to obtain apical four-chamber (A4C) and parasternal short-axis (PSAX) video clips, totaling 7,643 video clips and 17,600 annotated pictures. The video clips were partitioned into training (80%, n=6,114), testing (10%, n=765), and validation (10%, n=764) sets for machine learning purposes. In addition, 86% of the trials had an ejection fraction (EF) equal to or greater than 55%.

B. IMPLEMENTATION DETAILS

The computational setup consists of a single Tesla V100-32GB GPU, a 12-core CPU, and 61GB of RAM. The

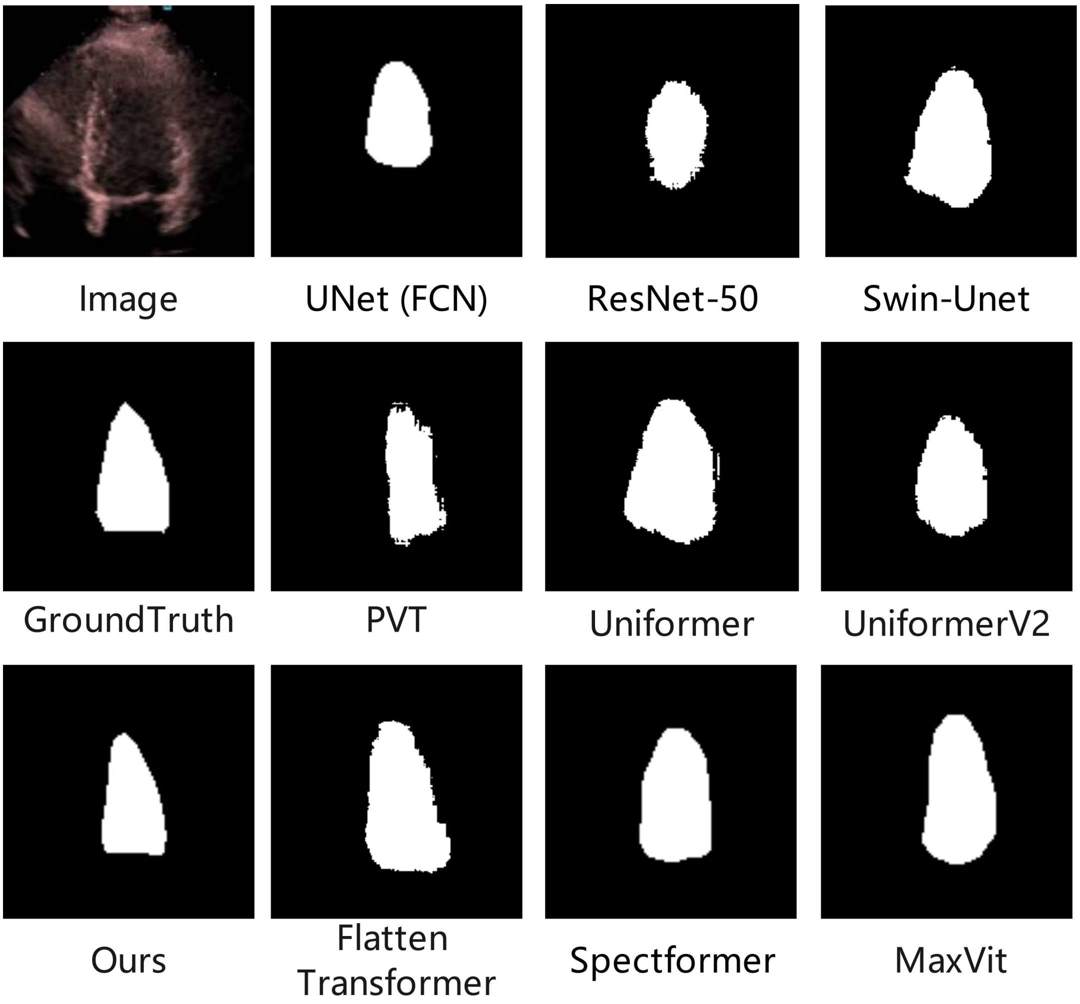


FIGURE 5. Visual comparisons of different methods on the A4C dataset.

system operates on an Ubuntu 18 environment with CUDA 11.0 and Pytorch 1.13 software.

The network was trained for 200 epochs, beginning with an initial learning rate of 1e-4. Batch sizes of 24 were selected for training to obtain a compromise between computational efficiency and model accuracy. The model's performance was assessed every five epochs, and early halting with a patience parameter of 10 was implemented to prevent overfitting. The network architecture was organized with layers configured in depth [2, 2, 2, 2]. The design depended on the multi-head attention mechanism with a specified number of heads [3, 6, 12, 24].

C. EVALUATION MEASURES

The current section demonstrates the statistical metrics adopted for comparing our predicted LV region (S_P) with the ground truth manually delineated by the experts (S_G). In this study, we employed three segmentation performance metrics: Dice similarity coefficient (DSC), Precision, and Recall [25]. The DSC measure is defined as the intersection of the PR and GT regions and is defined as:

$$DSC = \frac{2(S_P \cap S_G)}{S_P + S_G} \tag{5}$$

Precision and recall are determined by the number of true positives (TP), false positives (FP), and false negatives (FN) in the categorization and can be formulated in the following definitions:

$$Precision = (TP + FP)/TP \tag{6}$$
$$Recall = (TP + FN)/TP \tag{7}$$

where TP is the number of pixels or points accurately categorized as LV by both S_P and S_G and FP represents the pixels or points that S_P mistakenly classifies as LV but are not misclassified by S_G. FN indicates the pixels or points S_P misclassified as non-LV but are LV as defined by S_G.

To assess the computational efficiency of deep learning models and evaluate neural network performance, particularly for deployment in real-world applications or on specific

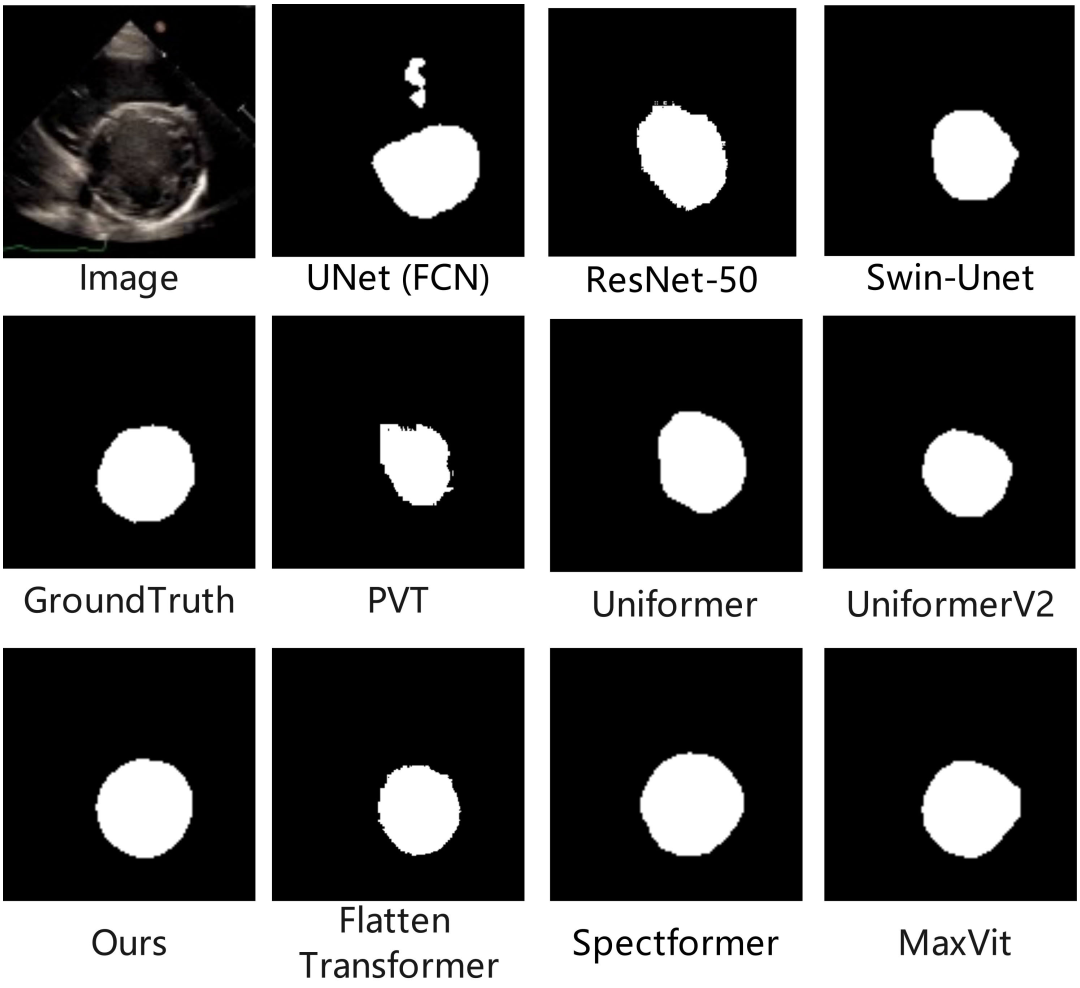


FIGURE 6. Visual comparisons of different methods on the PSAX dataset.

hardware, the following additional statistical measures are also used [26]:

1) Number of Parameters quantifies the combined number of parameters within the neural network model, which include weights and biases.
2) Inference Speed pertains to the speed at which the model can handle input data and generate output predictions during the inference phase.
3) GPU Memory represents the RAM capacity on the GPU needed to store the neural network model.
4) GFLOPs evaluate the computational efficiency of a neural network model based on the quantity of floating-point operations.

V. EVALUATION RESULTS AND DISCUSSION

This part presents the experimental findings to showcase the effectiveness of our method compared to other current methodologies. The study initially performed comparative experiments using the U-Net architecture with three different backbones [27]: FCN, DeepLabV3, and PSPNet, ResNet50 [28], in addition to other state-of-the-art segmentation models including Swin-Unet [7], Spectformer [22], Uniformer [30], [31], PVT [29], Flatten Transformer [32], and MaxVit [33]. The findings in **Table 2** demonstrate that our model outperforms other models regarding DSC on PSAX and A4C datasets, respectively, implying higher quality in shape-aware segmentation. The values in bold indicate the highest numbers for the respective criteria.

A. QUALITATIVE COMPARISON WITH OTHER METHODS

We provide a detailed qualitative visualization of the left ventricle segmentation results in **Figure 5** and **Figure 6**. Other approaches struggle with inaccurate segmentation of target forms in pediatric echocardiograms because they do not consider boundary information. Our proposed technique closely resembles the ground truths regarding shape similarity, which can offer more precise and detailed delineations of the left ventricle in pediatric echocardiograms.

B. ABLATION STUDIES

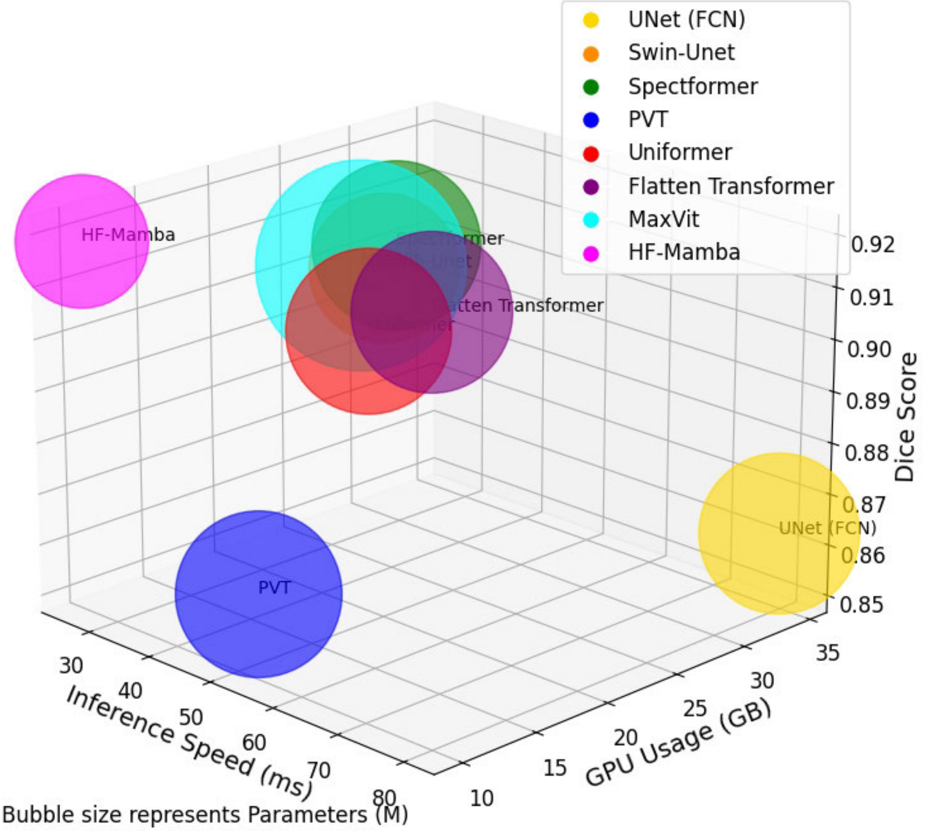
Initially, experiments were performed to explore the impact of changing the number of HFEB in the HF-Mamba encoder

TABLE 3. Ablation study of HFEB on performance metrics across PSAX and A4C datasets.

Methods	Dataset PSAX			Dataset A4C		
	Precision	Recall	DSC	Precision	Recall	DSC
HFE-Mamba-1	0.9146	0.9239	0.9099	0.9043	0.9061	0.8966
HFE-Mamba-2	0.9195	0.9289	0.9143	0.9090	0.9083	0.9010
HFE-Mamba-3	**0.9200**	**0.9290**	**0.9188**	**0.9149**	**0.9103**	**0.9029**

TABLE 4. Ablation study of vision mamba block on performance metrics across PSAX and A4C datasets.

Methods	Dataset PSAX			Dataset A4C		
	Precision	Recall	DSC	Precision	Recall	DSC
Mamba → PVT	0.9147	0.9240	0.9110	0.9043	0.9097	0.8976
Mamba → Flatten Transformer	0.9178	0.9272	0.9159	0.9072	0.9109	0.9005
Mamba → MaxVit	0.9197	0.9289	0.9167	0.9091	0.9101	0.9013
Ours	**0.9200**	**0.9290**	**0.9188**	**0.9149**	**0.9103**	**0.9029**



Bubble size represents Parameters (M)
Largest bubble (480.79M) is MaxVit

FIGURE 7. Visualization of efficiency comparisons between HFE-Mamba and other models on the PSAX dataset.

structure. We inserted 1, 2, and 3 HFEBs into the encoder path's first, second, and third stages, namely HFE-Mamba-1, HFE-Mamba-2, and HFE-Mamba-3, as shown in **Table 3**. The results demonstrate that including HFEBs leads to enhancements in both datasets in terms of evaluation metrics (DSC and Recall), highlighting the beneficial impact of our

TABLE 5. Comparison with other methods regarding parameters (M), inference speed (ms), GFLOPs, and GPU memory (GB) on the PSAX dataset.

Methods	#Params	Inference Speed	GPU memory	GFLOPs
UNet (FCN) [27]	279.41	80.52	34.50	284.60
Swin-Unet [7]	241.61	45.98	21.81	164.04
Spectformer [22]	308.24	47.39	22.11	178.63
PVT [29]	299.81	40.39	15.39	181.88
Uniformer [30]	297.24	54.93	16.84	181.01
Flatten Transformer [32]	282.57	67.48	15.55	144.45
MaxVit [33]	480.79	46.29	20.08	259.48
Ours	**192.13**	**25.85**	**9.89**	**92.64**

HFEB in enhancing shape-aware segmentation during feature extraction.

Table 4 examines the impact of replacing the Mamba branch with ViT structures of quadratic (PVT) and linear complexity (Flatten ViT, MaxViT). The results indicate that Vision Mamba may provide higher accuracy than other models with linear or quadratic complexity.

C. MODEL EFFICIENCY COMPARISON

Table 5 compares model efficiency between HP-Mamba, classic deep learning architectures, and ViT methods with quadratic and linear complexity based on model parameters, inference speed (ms), GPU memory usage, and GLOPs. Therefore, our P-Mamba exhibits superior performance across all metrics compared to other methods. It can be seen from **Figure 7** that our proposed method also has lower GFLOPs than state-of-the-art approaches, even though our segmentation performance surpasses them significantly. In addition, the attention-free architecture of our Mamba design significantly enhances model efficiency compared to linear complexity models.

VI. CONCLUSION

To conclude, this study introduces a novel deep network structure, HFE-Mamba, designed for segmenting the left ventricle in pediatric echocardiograms, specifically emphasizing high-frequency augmentation processes. HFE-Mamba successfully captures fine details and maintains a balance with the low-frequency domain by including HFEB in the encoding phase and applying cross-attention. HFE-Mamba utilizes vision mamba layers to enhance compute and memory efficiency through capturing global dependencies with a lightweight architecture with low computational requirements. Comprehensive tests on both LV datasets indicate that it performs well in various imaging scenarios.

REFERENCES

[1] D. Van Der Linde, E. E. Konings, M. A. Slager, M. Witsenburg, W. A. Helbing, J. J. Takkenberg, and J. W. Roos-Hesselink, "Birth prevalence of congenital heart disease worldwide: A systematic review and meta-analysis," *J. Amer. College Cardiol.*, vol. 58, no. 21, pp. 2241–2247, Nov. 2011.

[2] L. Xu, M. Liu, J. Zhang, and Y. He, "Convolutional-neural-network-based approach for segmentation of apical four-chamber view from fetal echocardiography," *IEEE Access*, vol. 8, pp. 80437–80446, 2020.

[3] J. F. Silva, J. M. Silva, A. Guerra, S. Matos, and C. Costa, "Ejection fraction classification in transthoracic echocardiography using a deep learning approach," in *Proc. IEEE 31st Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Karlstad, Sweden, Jun. 2018, pp. 123–128.

[4] R. M. Lang, L.P. Badano, V. Mor-Avi, J. Afilalo, A. Armstrong, L. Ernande, F. A. Flachskampf, E. Foster, S. A. Goldstein, and T. Kuznetsova, "Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging," *Eur. Heart J.-Cardiovascular Imag.*, vol. 16, no. 3, pp. 233–271, Mar. 2015.

[5] A. Power, S. Poonja, D. Disler, K. Myers, D. J. Patton, J. K. Mah, N. M. Fine, and S. C. Greenway, "Echocardiographic image quality deteriorates with age in children and young adults with Duchenne muscular dystrophy," *Frontiers Cardiovascular Med.*, vol. 4, p. 82, Dec. 2017.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[7] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, Oct. 2022, pp. 1–12.

[8] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.

[9] Z. Pan, J. Cai, and B. Zhuang, "Fast vision transformers with hilo attention," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Virtual Conf.*, Dec. 2022, pp. 1–11.

[10] Y. Zhong, B. Li, L. Tang, S. Kuang, S. Wu, and S. Ding, "Detecting camouflaged object in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 4494–4503.

[11] Y. Zeng, P.-H. Tsui, K. Pang, G. Bin, J. Li, K. Lv, X. Wu, S. Wu, and Z. Zhou, "MAEF-Net: Multi-attention efficient feature fusion network for left ventricular segmentation and quantitative analysis in two-dimensional echocardiography," *Ultrasonics*, vol. 127, Jan. 2023, Art. no. 106855.

[12] R. Naghne, A. Kazemi, H. Moghaddasi, M. Rahmani, P. Farnia, A. Ahmadian, and J. Alirezaie, "An efficient capsule-based network for 2D left ventricle segmentation in echocardiography images," in *Proc. 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2023, pp. 1–4.

[13] M. J. Mortada, S. Tomassini, H. Anbar, M. Morettini, L. Burattini, and A. Sbrollini, "Segmentation of anatomical structures of the left heart from echocardiographic images using deep learning," *Diagnostics*, vol. 13, no. 10, p. 1683, May 2023.

[14] H. Li, Y. Wang, M. Qu, P. Cao, C. Feng, and J. Yang, "EchoEFNet: Multi-task deep learning network for automatic calculation of left ventricular ejection fraction in 2D echocardiography," *Comput. Biol. Med.*, vol. 156, Apr. 2023, Art. no. 106705.

[15] M. Saeed, R. Muhtaseb, and M. Yaqub, "Contrastive pretraining for echocardiography segmentation with limited data," in *Proc. Annu. Conf. Med. Image Understand. Anal.* Cambridge, U.K., Jul. 2022, pp. 680–691.

[16] N. Painchaud, N. Duchateau, O. Bernard, and P.-M. Jodoin, "Echocardiography segmentation with enforced temporal consistency," *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2867–2878, Oct. 2022.

[17] L. Guo, B. Lei, W. Chen, J. Du, A. F. Frangi, J. Qin, C. Zhao, P. Shi, B. Xia, and T. Wang, "Dual attention enhancement feature fusion network for segmentation and quantitative analysis of paediatric echocardiography," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102042.

[18] Y. Hu, B. Xia, M. Mao, Z. Jin, J. Du, L. Guo, A. F. Frangi, B. Lei, and T. Wang, "AIDAN: An attention-guided dual-path network for pediatric echocardiography segmentation," *IEEE Access*, vol. 8, pp. 29176–29187, 2020.

[19] C. Zhao, B. Xia, W. Chen, L. Guo, J. Du, T. Wang, and B. Lei, "Multi-scale wavelet network algorithm for pediatric echocardiographic segmentation via hierarchical feature guided fusion," *Appl. Soft Comput.*, vol. 107, Aug. 2021, Art. no. 107386.

[20] C. D. Reddy, L. Lopez, D. Ouyang, J. Y. Zou, and B. He, "Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients," *J. Amer. Soc. Echocardiography*, vol. 36, no. 5, pp. 482–489, May 2023.

[21] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "PIDNet: A real-time semantic segmentation network inspired by PID controllers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19529–19539.

[22] B. N. Patro, V. P. Namboodiri, and V. S. Agneeswaran, "SpectFormer: Frequency and attention is what you need in a vision transformer," 2023, *arXiv:2304.06446*.

[23] B. N. Patro and V. S. Agneeswaran, "Scattering vision transformer: Spectral mixing matters," 2023, *arXiv:2311.01310*.

[24] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," 2024, *arXiv:2401.09417*.

[25] A. W. Setiawan, "Image segmentation metrics in skin lesion: Accuracy, sensitivity, specificity, dice coefficient, Jaccard index, and Matthews correlation coefficient," in *Proc. Int. Conf. Comput. Eng., Netw., Intell. Multimedia (CENIM)*, Nov. 2020, pp. 97–102.

[26] B. Paria, C.-K. Yeh, I. E. H. Yen, N. Xu, P. Ravikumar, and B. Póczos, "Minimizing FLOPs to learn efficient sparse representations," 2020, *arXiv:2004.05665*.

[27] D. Gupta, "Image segmentation Keras: Implementation of Segnet, FCN, UNet, PSPNet and other models in Keras," 2023, *arXiv:2307.13215*.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[29] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 548–558.

[30] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "UniFormer: Unifying convolution and self-attention for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12581–12600, Jun. 2023.

[31] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, L. Wang, and Y. Qiao, "UniFormerV2: Unlocking the potential of image ViTs for video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 1632–1643.

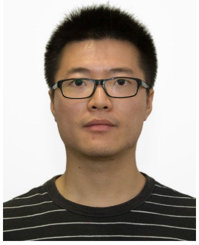
[32] D. Han, X. Pan, Y. Han, S. Song, and G. Huang, "FLatten transformer: Vision transformer using focused linear attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 5961–5971.

[33] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "MaxViT: Multi-axis vision transformer," in *Proc. 17th Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, 2022, pp. 459–479.

ZI YE received the master's degree in applied statistics from the University of Oxford, U.K., in 2010, and the Ph.D. degree from Universiti Teknikal Malaysia Melaka, in 2022. She is currently a Postdoctoral Researcher with the Institute of Intelligent Software, Guangzhou, China. Her research interests include artificial intelligence and machine learning.

TIANXIANG CHEN is currently pursuing the Ph.D. degree with the University of Science and Technology of China. His research interests include medical image analysis, small target detection, and multi-modality.

DAN WANG received the B.S. degree in software engineering from Nanchang University, China, in 2023. She is currently a Graduate Student in artificial intelligence with the University of Chinese Academy of Sciences. Her research interests include medical artificial intelligence and machine learning.



FANGYIJIE WANG received the dual master's degrees in computer science and statistics from University College Dublin, where he is currently pursuing the Ph.D. degree with the School of Medicine. He is a participant of the Science Foundation Ireland ML-Labs Program. His research interest includes AI applications in medical imaging with deep learning methodologies.

LIJUN ZHANG is the Director of the Sino-European Joint Laboratory for Reliable and Intelligent Software, Guangzhou Intelligent Software Industry Research Institute. His research interest includes model checking for probabilistic concurrent systems based on Markov models and their extensions.