FrameONE: Hierarchical Motion Modeling for Universal Multi-View Echocardiographic Keyframe Detection

Rusi Chen[1⋆], Yuhao Huang[1,2⋆], Hongyuan Zhang[2⋆], Chao Tian[3], Shunan Ji[4,5], Yuhan Zhang[1], and Dong Ni[1,3,4,5(✉)]

[1]Medical Ultrasound Image Computing (MUSIC) Lab, Shenzhen University, Shenzhen, China
nidong@szu.edu.cn
[2]Centre for Artificial Intelligence and Robotics (CAIR), Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hongkong, China
[3]School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China
[4]School of Artificial Intelligence, Shenzhen University, Shenzhen, China
[5]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China

Abstract. Accurate detection of end-systole (ES) and end-diastole (ED) frames is fundamental to echocardiographic assessment. Existing methods are typically developed in a view-specific manner, depend on auxiliary annotations or intensive visual modeling, which limits their generalizability. In multi-view modeling, keyframe detection is driven by shared cardiac motion, yet large appearance differences and motion patterns make unified modeling challenging. To address these issues, we propose **FrameONE**, a unified end-to-end framework for multi-view echocardiographic keyframe detection. FrameONE introduces a **Hierarchical Motion Modeling** strategy: an intra-view multi-task learning reduces appearance bias and promotes motion-focused representations within each view; an inter-view general motion learning module further separates view-agnostic dynamics from view-specific patterns, enabling shared yet flexible motion representation learning across views. Extensive experiments on 25,872 videos spanning four standard views demonstrate that FrameONE achieves state-of-the-art keyframe detection accuracy with strong cross-view generalization. Code is available at https://github.com/szuboy/FrameONE.

Keywords: Keyframe Detection · Multi-view Echocardiography · Hierarchical Motion Decomposition.

1 Introduction

The accurate identification of end-systole (ES) and end-diastole (ED) frames is fundamental to echocardiographic assessment, as clinical parameters are typi-

⋆ Rusi Chen, Yuhao Huang, and Hongyuan Zhang contributed equally.

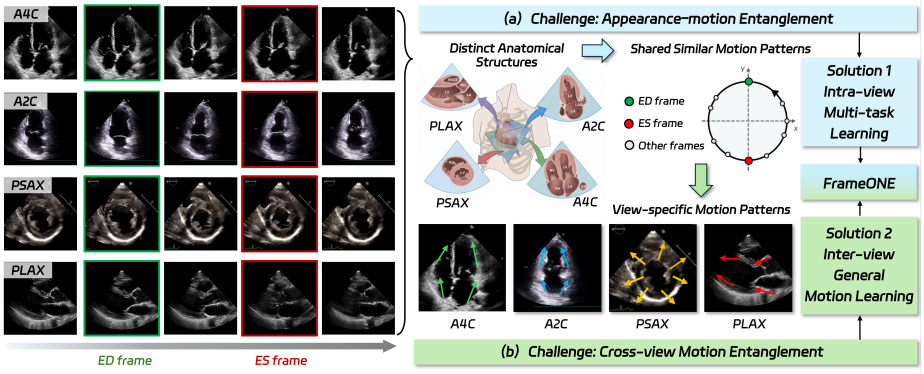

Fig. 1. Overview of the two core challenges in multi-view cardiac keyframe detection and the corresponding solutions in FrameONE. Four echocardiographic views are: apical four-chamber (A4C), apical two-chamber (A2C), parasternal long-axis (PLAX), and parasternal short-axis (PSAX).

cally derived from these keyframes across multiple standardized views [8]. However, the anatomical features defining these keyframes vary significantly among different views, hindering consistent and reliable identification (Fig. 1a) [12]. Besides, manual assessment is labor-intensive and susceptible to intra- and inter-observer variability. Therefore, these challenges motivate the development of view-agnostic automated keyframe detection model.

Existing deep learning approaches have shown satisfactory progress in intelligent heart ultrasound analysis [18,21,2,19,20]. However, for ES/ED frame detection, most methods are predominantly designed for single-view settings and often rely on view-specific assumptions [7,4,16,1,9]. Several supervised methods incorporate additional anatomical or clinical annotations. For instance, Feng et al. [5] jointly detect keyframes and landmarks, but rely on view-specific landmark motion that may limit cross-view generalization. Similarly, Reynaud et al. [15] adopt a multi-task framework that requires extra ejection fraction supervision, increasing annotation demands. Yang et al. [17] introduced an unsupervised decoupling method, but it is strictly tailored to the A4C view by assuming orthogonal motion directions specific to A4C anatomy. Furthermore, Lu et al. [11] integrate optical flow to explicitly model cardiac dynamics, at the expense of increased computational overhead and sensitivity to redundant visual information. Overall, previous methods remain largely view-dependent and do not explicitly leverage shared physiological patterns or complementary information across multiple views, limiting their clinical applicability.

In this work, instead of view-specific designs, we propose an end-to-end unified framework for detecting ES and ED frames across multiple echocardiographic views, named **FrameONE**. As shown in Fig. 1, developing such a unified model faces two fundamental challenges. 1) Although different views share the

same cardiac rhythm, they exhibit distinct anatomical structures. Consequently, models may tend to focus on view-specific appearance features rather than motion dynamics. 2) Even under the same cardiac cycle, motion patterns differ across views, such as longitudinal contraction in A4C and radial contraction in PSAX. Overall, the main contributions of this work are as follows:

- **A pioneering unified framework**. We present FrameONE, to the best of our knowledge, is the first attempt to explore multi-view echocardiographic keyframe detection without view-specific design.
- We introduce **Hierarchical Motion Modeling (HMM)**, including Intra-view Multi-task Learning (IML) and an Inter-view General Motion Learning (IGM) strategies. IML disentangles the representation of motion and structural features, IGM decouples motion into view-agnostic and view-specific patterns, enabling shared motion representation learning across views.
- Extensive experiments across four echocardiographic views demonstrate that FrameONE outperforms existing methods on most evaluation metrics.

2 Method

As illustrated in Fig. 2, given an echocardiographic video $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ from arbitrary view (A4C/A2C/PLAX/PSAX), FrameONE detects ES and ED frames by regressing a continuous score for each frame, where 0 corresponds to ES and 1 to ED. The framework employs a shared encoder to extract inter-view motion features, followed by an IGM module that disentangles view-agnostic and view-specific representations. These features are then fed into a dual-decoder for ES/ED phase regression and view classification.

2.1 Intra-view Multi-task Learning

Different echocardiographic views have distinct appearances, which can make the model rely on visual cues rather than cardiac motion. To address this issue, we introduce IML to jointly optimize keyframe regression and view classification by a dual-decoder learning paradigm. This design encourages the network to focus on motion patterns that are consistent for keyframe detection.

As illustrated in Fig. 2a, each frame $\mathbf{x}_t$ is encoded by a shared ResNet18 [6] backbone with a fully connected layer $\mathbf{f}_t = \mathcal{E}(\mathbf{x}_t) \in \mathbb{R}^{256}$. However, static frame features alone cannot capture the dynamic contraction-relaxation process essential for identifying ES/ED. While optical flow for model motion may introduce computational overhead and is sensitive to speckle noise and probe motion. Here, we extract short-term motion representations by applying a learnable 1D temporal convolution over the feature sequence $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_T]^\top \in \mathbb{R}^{T \times d}$:

$$\Delta\mathbf{f}_t = \text{Conv1D}(\mathbf{F}_t), \quad t = 1, \ldots, T-1, \tag{1}$$

where the convolution operates along the temporal axis with kernel size k. This formulation aligns naturally with the task: ES and ED correspond to motion extrema (sign changes in $\Delta\mathbf{f}_t$) rather than appearance extrema.

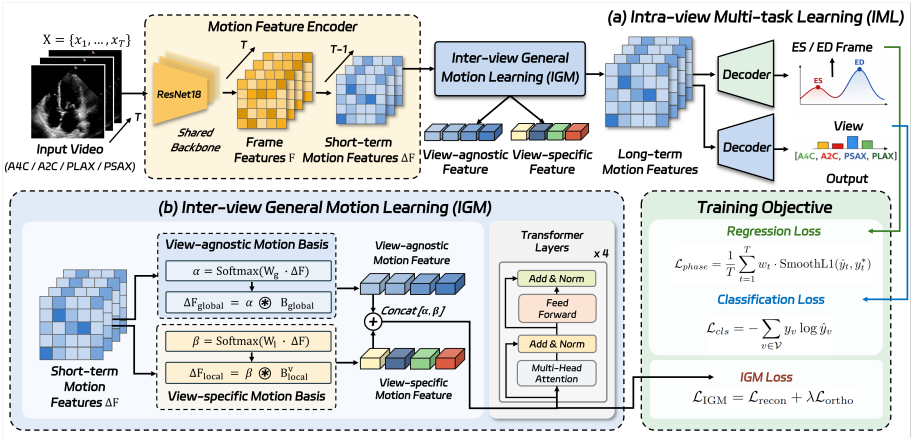

Fig. 2. Overview of proposed FrameONE.

The multi-task decoder consists of two heads: (i) a regression head that outputs frame-level scores $\hat{y}_t \in [0, 1]$, and (ii) a view classification head that predicts the imaging view. The regression loss is a weighted Smooth L1 loss:

$$\mathcal{L}_{phase} = \frac{1}{T} \sum_{t=1}^{T} w_t \cdot \text{SmoothL1}(\hat{y}_t, y_t^*), \tag{2}$$

where the per-frame weights w_t assign higher importance to frames near the ES/ED boundaries. To obtain w_t, we construct a Gaussian keyframe-proximity map $w_{\text{gauss}}(t)$ by centering 1D Gaussian kernels with $\sigma = 2$ at the ES and ED frame indices and aggregating their responses over time. The final weighting is defined as $w_t = 1 + w_{\text{gauss}}(t)$, which up-weights the keyframe neighborhoods and encourages the model to prioritize regression accuracy around ES/ED. The view classification loss is defined as the standard cross-entropy:

$$\mathcal{L}_{cls} = -\sum_{v \in \mathcal{V}} y_v \log \hat{y}_v, \tag{3}$$

where $\mathcal{V}$ denotes the set of view categories, $y_v \in \{0, 1\}$ is the one-hot ground-truth indicator for class v, and $\hat{y}_v \in [0, 1]$ is the predicted probability for view v produced by the classification head.

2.2 Inter-view General Motion Learning

Cardiac views share a common contraction-relaxation rhythm that reflects global cardiac physiology. However, each view also presents its own motion characteristics (see Fig. 1b). Yet, learning motion features directly from multi-view data often mix view-specific motion with the global contraction–relaxation pattern, which obscures the underlying signal and leads to suboptimal results.

To address this issue, we disentangle motion into a shared rhythm component and a view-dependent residual component (see Fig. 2b). The core idea is to factorize representations into two spaces: a view-agnostic global subspace and a view-specific local subspace. Specifically, let $\mathbf{B}_{global}$ be a global basis matrix whose K_g columns capture universal cardiac dynamics shared across all views, and $\mathbf{B}^v_{local}$ be a view-specific basis matrix whose K_l columns model local motion variations for view v. The reconstruction motion is then decomposed as:

$$\Delta\hat{\mathbf{f}}_t = \Delta\hat{\mathbf{f}}^g_t + \Delta\hat{\mathbf{f}}^l_t = \boldsymbol{\alpha} \cdot \mathbf{B}_{global} + \boldsymbol{\beta} \cdot \mathbf{B}^v_{local}, \tag{4}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ represent projection coefficients in the global and local subspaces. We then apply a softmax function to encourage sparse basis activation:

$$\boldsymbol{\alpha} = \mathrm{softmax}(\mathbf{W}_g^\top \Delta\mathbf{f}_t), \quad \boldsymbol{\beta} = \mathrm{softmax}(\mathbf{W}_l^\top \Delta\mathbf{f}_t), \tag{5}$$

where $\mathbf{W}_g \in \mathbb{R}^{d\times K_g}$ and $\mathbf{W}_l \in \mathbb{R}^{d\times K_l}$ are learnable projection matrices. This formulation encourages the model to separate shared physiological dynamics from view-dependent motion patterns. Then, the global and local coefficient vectors are concatenated as $\mathbf{z}_t = \mathrm{Concat}(\boldsymbol{\alpha}; \boldsymbol{\beta})$, forming a compact motion descriptor at each time step. Over a cardiac cycle, the sequence $\mathbf{z}_t$ traces a low-dimensional trajectory that characterizes contraction and relaxation dynamics. Since ES and ED correspond to extrema of this cyclic process rather than local motion magnitude, we model the global temporal structure of this trajectory using a 4-layer transformer temporal encoder, yielding contextualized representations. Then, they are fed to the multi-task decoder to regress scores and classify view types.

Finally, in order to ensure proper disentanglement between shared and view-specific dynamics, we adopt a reconstruction objective $\mathcal{L}_{\text{recon}}$ together with a cross-subspace orthogonality constraint $\mathcal{L}_{\text{ortho}}$ to formulate the loss function as:

$$\begin{aligned}\mathcal{L}_{\text{IGM}} &= \mathcal{L}_{\text{recon}} + \lambda\mathcal{L}_{\text{ortho}} \\ &= \frac{1}{T-1}\sum_{t=1}^{T-1} \|\Delta\hat{\mathbf{f}}_t - \Delta\mathbf{f}_t\|_2^2 + \lambda\|\mathbf{U}_g^\top \mathbf{U}_l^v\|_F^2 + \lambda\|\mathbf{z}_t\|_1,\end{aligned} \tag{6}$$

where λ controls the orthogonality constraints ($\lambda = 0.1$ based on empirical observations). $\mathcal{L}_{\text{IGM}}$ mitigates leakage between global rhythm and view-dependent components, which improves identifiability and stabilizes training.

2.3 Training Objective and Inference

Training. The overall FrameONE is trained in an end-to-end manner by jointly optimizing the phase regression objective and multiple auxiliary regularization terms. The overall loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{phase} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{\text{IGM}}\mathcal{L}_{\text{IGM}}, \tag{7}$$

where $\lambda_{cls} = \lambda_{\text{IGM}} = 0.1$. The supervision signal is provided by $\mathcal{L}_{phase}$, while the auxiliary view classification loss $\mathcal{L}_{cls}$ guides the shared encoder to preserve view-discriminative anatomical information. In addition, the IGM regularization term further ensures more stable optimization.

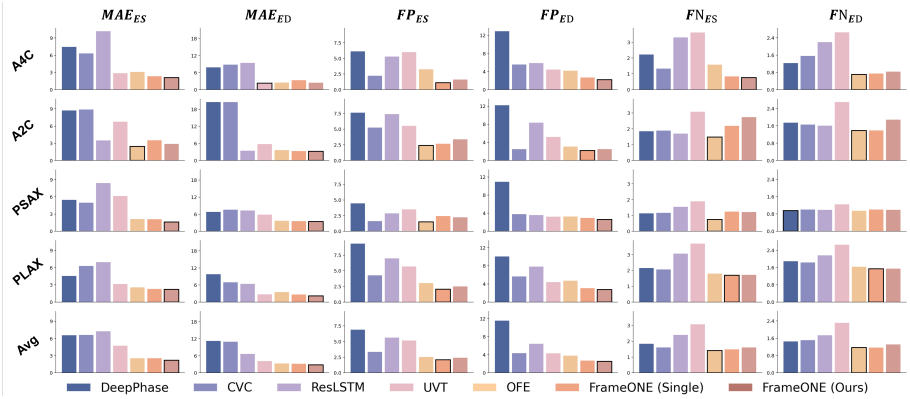

Fig. 3. Quantitative comparison of ES/ED detection across four views.

Inference. During testing, we use a sliding-window strategy with 50% overlap to process videos of arbitrary length and average predictions in overlapping regions. The score curve is thresholded at 0.5 to obtain candidate segments. Consecutive candidates are grouped, and the peak frame within each group is selected as the local extremum using a 5-frame window. If no candidate is detected, the global minimum and maximum of the score curve are used as fallback predictions.

3 Experiments and Results

Implementation Details. Our model was implemented with PyTorch on two NVIDIA GeForce RTX 4090 GPUs. All input frames were resampled to 128×128. We followed [15] to duplicate the transition frames between the two labeled frames and append them after the last annotation. During training, we employed a random start fixed-length sampling strategy to extract 50 frames from each video. AdamW optimizer [10] was used with an initial learning rate of 5e-3 and weight decay of 1e-5. We used a cosine annealing learning rate scheduler with 5 warmup epochs. The batch size was 32, and the total training epoch was 100.

Datasets and Evaluation Metrics. We evaluate our method on a large cross-population echocardiographic cohort covering four cardiac views. The public part includes A4C from EchoNet-Dynamic [13], PLAX from EchoNet-LVH [3], and PSAX from Echo-pediatric [14], while the A2C view is from our private dataset. In total, we collected 25,872 samples, comprising 9,926 A4C, 11,021 PLAX, 4,475 PSAX, and 450 A2C videos. The overall videos are split into 20,338, 3,367, and 2,167 for training, validation, and testing, respectively. Since public datasets provide only one ES/ED pair label per video. To evaluated method performance in read-world clinical settings, we additionally annotated 50–100 videos per view with ES/ED labels across 3-5 cardiac cycles by experienced sonographers. Thus,

Table 1. Comparison results with other state-of-the-art keyframe detection methods across four views. FrameOne(Single) means model trained with single view data. The best results are colored in blue, and the second are underlined.

View	Method	Single-cycle			Multi-cycle	
		MAE ↓	RMSE↓	R^2↑	FP↓	FN↓
A4C	DeepPhase [4](SR, 2023)	7.707	24.581	0.619	9.556	1.737
	CVC [16](AIM, 2023)	7.606	18.566	0.783	3.917	1.453
	ResLSTM [7](CBM, 2021)	9.844	12.115	0.907	5.613	2.762
	UVT [15](MICCAI, 2021)	2.599	17.450	0.867	5.242	3.144
	OFM [11](TMI, 2025)	2.832	3.527	0.994	3.762	1.149
	FrameONE (Single)	2.883	3.980	0.993	1.925	0.803
	FrameONE (Ours)	2.339	2.974	0.995	1.920	0.803
A2C	DeepPhase [4](SR, 2023)	14.679	28.569	0.286	9.952	1.809
	CVC [16](AIM, 2023)	14.771	27.360	0.342	3.904	1.785
	ResLSTM [7](CBM, 2021)	3.534	4.340	0.985	7.904	1.666
	UVT [15](MICCAI, 2021)	6.302	12.817	0.893	5.404	2.904
	OFM [11](TMI, 2025)	3.077	3.972	0.989	2.738	1.428
	FrameONE (Single)	3.464	6.339	0.972	2.450	1.800
	FrameONE (Ours)	3.121	4.854	0.984	2.974	2.325
PSAX	DeepPhase [4](SR, 2023)	6.204	17.589	0.782	7.731	1.052
	CVC [16](AIM, 2023)	6.347	13.202	0.871	2.731	1.097
	ResLSTM [7](CBM, 2021)	7.887	10.664	0.920	3.253	1.276
	UVT [15](MICCAI, 2021)	6.032	15.193	0.827	3.388	1.582
	OFM [11](TMI, 2025)	2.970	5.136	0.978	2.395	0.850
	FrameONE (Single)	2.891	5.274	0.979	2.719	1.136
	FrameONE (Ours)	2.520	6.171	0.971	2.424	1.113
PLAX	DeepPhase [4](SR, 2023)	7.204	20.828	0.678	9.740	2.030
	CVC [16](AIM, 2023)	6.656	13.947	0.861	5.010	1.955
	ResLSTM [7](CBM, 2021)	6.722	8.934	0.943	7.450	2.625
	UVT [15](MICCAI, 2021)	2.992	17.339	0.932	5.090	3.185
	OFM [11](TMI, 2025)	3.071	5.026	0.993	3.920	1.725
	FrameONE (Single)	2.509	4.150	0.996	2.573	1.615
	FrameONE (Ours)	2.250	4.376	0.995	2.647	1.647
Avg	OFM [11](TMI, 2025)	2.907	4.151	0.990	3.438	1.173
	FrameONE (Ours)	2.386	3.994	0.989	2.176	1.042

we evaluate the model under *single-cycle annotation* and *multi-cycle annotation* settings. The former uses MAE, RMSE, and R^2 as metrics, while the latter assesses FP and FN rates, defined as $N_{\text{FP}}/N_{\text{pred}}$ and $N_{\text{FN}}/N_{\text{GT}}$, respectively, where N_{FP} and N_{FN} denote unmatched predictions and missed ground truths. All metrics are reported at the frame level.

Comparison Study. As shown in Table 1, FrameONE consistently achieves the best performance across all four echocardiographic views. Compared with the state-of-the-art OFM, FrameONE further reduces the overall MAE from 2.907

Table 2. Ablation study results.

IML	IGM		MAE↓
	Global Basis	Local Basis	
×	×	×	3.093
✓	×	×	2.700
✓	✓	×	2.470
✓	×	✓	2.480
✓	✓	✓	**2.362**

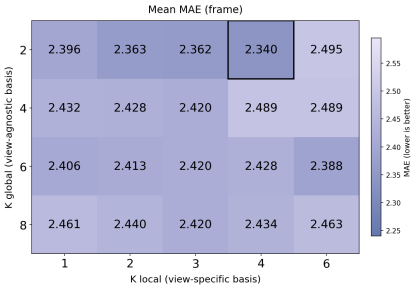

Fig. 4. Effect of basis number K_{g} and K_l in the IGM module.

to 2.386 frames ($p<0.001$) and decreases FP from 3.438 to 2.176 ($p<0.001$). In addition, FrameONE achieves a mean MAE of 2.339 frames on A4C, yielding 17.4% and 10.0% lower errors than OFM and UVT, respectively. Similar improvements are observed on the challenging PSAX and PLAX with less distinctive anatomies. In Fig. 3, we further report detailed ES and ED metrics, demonstrating that FrameONE achieves the best performance in most cases. For the competing methods, DeepPhase and CVC underperform due to the missing temporal contexts of CNNs. UVT lacks explicit motion decomposition and depends on auxiliary ejection fraction supervision. ResLSTM and OFM rely on complex temporal modeling (e.g., optical flow), incurring high computational cost. Moreover, FrameONE achieves 246.6 frame per second (FPS) with only 14.35M parameters, outperforming OFM (132.1 FPS, 43.93M) and ResLSTM (49.2 FPS, 29.81M). This further confirms the effectiveness of our method.

The multi-label testing set supports robustness evaluation, where FrameONE attains substantially lower FP (1.920 *vs.* 3.762/5.242) and FN (0.803 *vs.* 1.149/3.144) than the strong OFM and UVT (see Table 1). It is noted that FrameONE (Single) trained independently per view already outperforms all baselines, while our unified multi-view model further reduces MAE by ∼0.3-0.5 across views, demonstrating the necessity of our designed joint learning.

Ablation Study. Table 2 comprehensively tests each component of HMM, starting from a baseline model that directly predicts keyframes from raw frame features. First, adding IML reduces mean MAE from 3.093 to 2.700 frames, validating the importance of explicit inter-view learning. Incorporating the global basis further reduces MAE to 2.470 frames, demonstrating that capturing universal cardiac rhythm patterns across views benefits keyframe detection. Besides, the local basis yields a 0.613-frame reduction in MAE, suggesting the importance of view-dependent motion modeling. With all components, FrameONE achieves the best performance, validating the synergistic benefits of our hierarchical design. In Fig. 4, we carefully analyze the model sensitivity to the basis number. The mean MAE varies only slightly (2.340–2.495) across different settings, indicating that extensive parameter tuning is not required for FrameOne.

4 Conclusion

In this work, we presented FrameONE, the first unified framework for multi-view echocardiographic keyframe detection without view-specific design. The key idea of proposed method is hierarchical motion modeling: IML disentangles motion representations from view-dependent appearance within each view, while IGM further factorizes the learned motion into shared cardiac rhythms and view-specific subspaces, forming a tightly coupled hierarchy. Extensive experiments across four echocardiographic views demonstrate state-of-the-art performance. Future work will explore extending FrameONE to additional cardiac views and integrating it with downstream functional assessment tasks.

Acknowledgments. This work was supported by the grant from National Natural Science Foundation of China (12326619, 62572324); Frontier Technology Development Program of Jiangsu Province (BF2024078); and Science and Technology Planning Project of Guangdong Province (2023A0505020002).

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

References

1. Chen, H., Li, Y., Yang, L., Wu, H., Zhou, L., Sun, K., Shen, D.: A semi-supervised knowledge distillation framework for left ventricle segmentation and landmark detection in echocardiograms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 34–43. Springer (2025)
2. Chen, R., Yang, Y., Yao, J., Song, H., Zhang, J., Zhou, Y., Huang, Y., et al.: Mtcnet: Motion and topology consistency guided learning for mitral valve segmentation in 4d ultrasound. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 407–417. Springer (2025)
3. Duffy, G., Cheng, P.P., Yuan, N., He, B., Kwan, A.C., Shun-Shin, M.J., Alexander, K.M., Ebinger, J., Lungren, M.P., Rader, F., et al.: High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. JAMA cardiology **7**(4), 386–395 (2022)
4. Farhad, M., Masud, M.M., Beg, A., Ahmad, A., Ahmed, L.A., Memon, S.: Cardiac phase detection in echocardiography using convolutional neural networks. Scientific reports **13**(1), 8908 (2023)
5. Feng, Y., Yang, J., Li, M., Tang, L., Sun, S., Wang, Y.: A bayesian network for simultaneous keyframe and landmark detection in ultrasonic cine. Medical Image Analysis **97**, 103228 (2024)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Lane, E.S., Azarmehr, N., Jevsikov, J., Howard, J.P., Shun-Shin, M.J., Cole, G.D., Francis, D.P., Zolgharni, M.: Multibeat echocardiographic phase detection using deep neural networks. Computers in Biology and Medicine **133**, 104373 (2021)

8. Lang, R.M., Badano, L.P., Mor-Avi, V., Afilalo, J., Armstrong, A., Ernande, L., Flachskampf, F.A., Foster, E., Goldstein, S.A., Kuznetsova, T., et al.: Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging. European Heart Journal-Cardiovascular Imaging **16**(3), 233–271 (2015)
9. Li, H., Wang, Y., Qu, M., Cao, P., Feng, C., Yang, J.: Echoefnet: Multi-task deep learning network for automatic calculation of left ventricular ejection fraction in 2d echocardiography. Computers in Biology and Medicine **156**, 106705 (2023)
10. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019), https://arxiv.org/abs/1711.05101
11. Lu, Y., Tan, G., Pu, B., Yeung, P.H., Wang, H., Li, S., Rajapakse, J.C., Li, K.: Optical flow-enhanced mamba u-net for cardiac phase detection in ultrasound videos. IEEE Transactions on Medical Imaging (2025)
12. McLeod, K., Sermesant, M., et al.: Spatio-temporal tensor decomposition of a polyaffine motion model for a better analysis of pathological left ventricular dynamics. IEEE transactions on medical imaging **34**(7), 1562–1575 (2015)
13. Ouyang, D., He, B., Ghorbani, A., Lungren, M.P., et al.: Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. In: NeurIPS ML4H Workshop: Vancouver, BC, Canada. vol. 5, p. 2 (2019)
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. Reynaud, H., Vlontzos, A., Hou, B., et al.: Ultrasound video transformers for cardiac ejection fraction estimation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 495–505. Springer (2021)
16. Tasken, A.A., Berg, E.A.R., Grenne, B., Holte, E., Dalen, H., Stølen, S., Lindseth, F., Aakhus, S., Kiss, G.: Automated estimation of mitral annular plane systolic excursion by artificial intelligence from 3d ultrasound recordings. Artificial Intelligence in Medicine **144**, 102646 (2023)
17. Yang, Y., Yang, Q., Cui, K., Peng, C., D'Alberti, E., Hernandez-Cruz, N., Patey, O., Papageorghiou, A.T., Noble, J.A.: Latent motion profiling for annotation-free cardiac phase detection in adult and fetal echocardiography videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 316–325. Springer (2025)
18. Yu, J., Chen, R., Zhou, Y., Chen, Y., Duan, Y., Huang, Y., Zhou, H., Tan, T., Yang, X., Ni, D.: Explainable and controllable motion curve guided cardiac ultrasound video generation. In: International workshop on machine learning in medical imaging. pp. 232–241. Springer (2024)
19. Zhou, H., Chen, R., Yang, X., Chang, A., Yu, J., Huang, Y., Huang, R., Zhou, X., Zhou, L., Liang, J., et al.: Onuvs: An online motion transfer framework with content-texture decoupling for high-fidelity ultrasound video synthesis. IEEE Journal of Biomedical and Health Informatics (2026)
20. Zhou, X., Huang, Y., Dou, H., Chen, S., Chang, A., Liu, J., Long, W., Zheng, J., Xu, E., Ren, J., et al.: Ctrl-genaug: Controllable generative augmentation for medical sequence classification. International Journal of Computer Vision **134**(5), 216 (2026)
21. Zhou, X., Huang, Y., Xue, W., Dou, H., Cheng, J., Zhou, H., Ni, D.: Heartbeat: towards controllable echocardiography video synthesis with multimodal conditions-guided diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 361–371. Springer (2024)