Echo4DIR: 4D Implicit Heart Reconstruction from 2D Echocardiography Videos

Yanan Liu[1,2⋆], Qinya Li[1], Hao Zhang[2], Kangjian He[2], Xuan Yang[1], Hao Li[1], Dan Xu[2], and Lei Li[1✉]

[1] Department of Biomedical Engineering, National University of Singapore, Singapore
lei.li@nus.edu.sg

[2] School of Information Science and Engineering, Yunnan University, Kunming, China

Abstract. Reconstructing 4D (3D+t) cardiac geometry from sparse 2D echocardiography is highly desirable yet fundamentally challenged by geometric ambiguity and temporal discontinuity. To tackle these issues, we propose Echo4DIR, a novel test-time 4D implicit reconstruction framework. Specifically, we learn robust 3D shape priors from statistical shape models (SSMs) via a cardiac conditional SDF, constructing an Epipolar Mask Encoder module with epipolar cross attention to effectively fuse multi-view features. To bridge the synthetic-to-real domain gap, we introduce a self-supervised SDF-tailored differentiable rendering strategy for patient-specific 3D shape adaptation using uncalibrated clinical masks without requiring 3D ground truth. Crucially, the inherent continuity of implicit representation overcomes sparse observations, enabling anatomically reliable geometry at arbitrary resolutions. Furthermore, to empower our framework with physically continuous 4D extension, we introduce a Radial SDF Alignment strategy that strictly locks shape evolution to the predicted velocity field, fundamentally eliminating mesh drift. Extensive experiments on synthetic benchmarks and real clinical datasets demonstrate that Echo4DIR achieves state-of-the-art 4D cardiac mesh reconstruction, notably yielding an impressive clinical overlap of up to 98.35% Dice and 96.75% IoU. The code will be available in https://github.com/ryannus2025-ai/Echo4DIR.

Keywords: 4D Cardiac Reconstruction · Echocardiography · Implicit Neural Representation · Test-Time Optimization.

1 Introduction

Patient-specific 3D+t cardiac anatomical twins [11,21], which enable the automated assessment of cardiac morphology [6], motion [22], and function [7], hold immense promise for the early diagnosis and surgical navigation of cardiovascular diseases. While cardiac Computed Tomography (CT) [1], and Magnetic Resonance Imaging (MRI) [22], provide high-quality static anatomical details, their

⋆ completed during the YNU-NUS CSC joint Ph.D. program.

high cost or potential radiation exposure restrict their application in frequent examinations. In contrast, 2D Ultrasound (US) [9], despite being ubiquitous and real-time, inherently lacks 3D spatial context and relies heavily on clinician experience, introducing significant diagnostic uncertainty. Furthermore, 3D US [2] is often compromised by motion artifacts. Therefore, reconstructing 3D+t cardiac structures from multi-view 2D US emerges as a promising approach to achieve low-cost, high-precision, and personalized cardiac anatomical twins.

Recently, several impressed approaches have emerged. Stojanovsk et al. [18] reconstructed 3D cardiac voxels from multi-view synthetic 2D US, which ignores the synthetic-real domain shift. Laumer et al. [9] proposed a weakly supervised single-view reconstruction framework, training an image-mesh autoencoder [23] using cycle-consistency loss. Subsequently, Li et al. [21] introduced a generative framework based on diffusion models [17], learning the conditional distribution of 3D cardiac anatomy in a latent space with the CT supervision. Yet, the probabilistic nature of such models may lead to anatomical hallucinations, hindering the patient-specific reconstruction. A recent approach [10] integrated Graph Convolutional Networks (GCNs) [20] for 3D+t left ventricular (LV) mesh reconstruction, trying to capture the diversity of clinical cases via explicit differentiable rendering. Despite these advancements, reconstructing 3D+t cardiac geometry from sparse 2D US views remains the challenges: **1)** *Geometric Ambiguity.* The inherent sparsity of 2D echocardiography views induces significant information loss, particularly in unseen areas, making the model difficult to determine the underlying 3D shape. 2) *Temporal Discontinuity.* Directly extending static reconstruction to 3D+t sequences fails to establish dense anatomical correspondences across frames. This lack of explicit kinematic constraints leads to temporal artifacts such as mesh flickering and non-physical deformations.

To tackle these challenges, we propose **Echo4DIR**, a pioneering **Echo**cardiography based **4D I**mplicit cardiac **R**econstruction framework via conditional Signed Distance Functions (SDFs) [16]. Specifically, Echo4DIR integrates three key designs: **1)** We pioneer the integration of conditional SDFs into echo-based 3D+t cardiac shape reconstruction to optimize geometric ambiguity. Utilizing multi-view 2D mask features fused by the Epipolar Cross Attention (ECA) [19] as the condition, we learn a continuous 3D cardiac SDF trained by SSM priors. **2)** We introduce a Differentiable Rendering (DR) strategy tailored for SDFs to bridge the domain gap between synthetic SSMs and real clinical data, which fine-tunes the implicit surface using real echo masks and learnable transducer parameters, yielding high-fidelity, patient-specific cardiac geometry. **3)** We introduce a Radial SDF Alignment (RSA) strategy driven by the learned velocity field to harmonize spatial shape with temporal motion correlations, which resolves the topological inconsistency (mesh drift and volume collapse) inherent in SDF-derived meshes, achieving effective 3D+t dynamic reconstruction. Essentially, our proposed Echo4DIR operates as a Test-Time Optimization (TTO) framework enabling patient-specific shape adaptation. By exploiting the continuity of SDFs, it enables the plausible completion of unobserved geometry at arbitrary resolutions and robustly extends SDFs to the dynamic 3D+t domain.

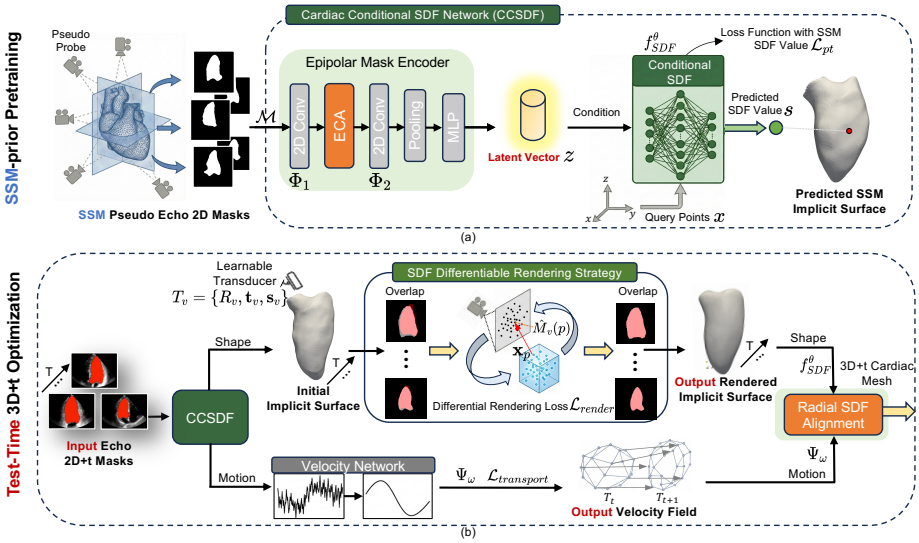

Fig. 1: Overall architecture of our proposed Echo4DIR framework.

2 Methodology

Overall Architecture. As shown in Fig. 1, we propose Echo4DIR, a pioneering 3D+t cardiac reconstruction framework via conditional SDFs from sparse-view 2D echo masks, consisting of: **1)** SSM-pretrained Cardiac Conditional SDF network (CCSDF) to learn cardiac shape priors. As shown in Fig. 1 (a), we simulate the US probe imaging on SSM meshes to generate pseudo masks from different views, which are encoded by the Epipolar Mask Encoder (EME) module integrated with a Epipolar Cross Attention (ECA) into the latent vector as the condition of SDFs. **2)** Test-time 3D+t Reconstruction Optimization. As shown in Fig. 1 (b), for shape optimization, clinical 2D echo masks are fed into CCSDF with learnable transducer to initialize the implicit surface. We introduce a SDF-tailored differentiable rendering strategy to enable self-supervised shape adaptation. For motion optimization, we introduce the Velocity Network to predict an implicit velocity field for any SDF point over time, integrating with a Radial SDF Alignment strategy to lock shape evolution with the velocity field, further achieving the robust 3D+t reconstruction. Below, we describe the details.

2.1 3D Cardiac Shape Representation via Conditional Signed Distance Function

To overcome the discretization artifacts in voxel and mesh based representations, we introduce the deep-learning conditional Signed Distance Function (SDF) [16]. Firstly, we define the SDF as a continuous mapping $SDF(\cdot) : \mathbb{R}^3 \rightarrow \mathbb{R}$. For any spatial query point $\mathbf{x} \in \mathbb{R}^3$, the value $s = SDF(\mathbf{x}) \in \mathbb{R}$ represents the Euclidean

distance to the nearest surface boundary $\mathcal{S}$, where the sign indicates the interior or exterior regions. The cardiac surface is implicitly represented by the zero-level set $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^3 \mid s = SDF(\mathbf{x}) = 0\}$. Next, we encode the individual target shape into a latent vector $\mathbf{z}$ and parameterize $SDF(\cdot)$ using a Multi-Layer Perceptron (MLP) [4]. The conditional SDF f^{θ}_{SDF} can be formulated as:

$$f^{\theta}_{SDF}(\mathbf{z}, \mathbf{x}) \approx SDF(\mathbf{x}), \forall \mathbf{x} \in \Omega, \tag{1}$$

where θ denotes the learnable parameters of the MLP, $\mathbf{z} \in \mathbb{R}^L$ is the latent vector encapsulating the patient-specific geometry and L is the dimension of the latent space. $\Omega \subset \mathbb{R}^3$ represents the set of spatial query points.

Epipolar Mask Encoder. We construct an Epipolar Mask Encoder to generate a personalized latent vector from v-view echo masks. First, a shared-weight 2D convolutional backbone [5] Φ_1 is used to extract the feature map $F_v = \Phi_v(M_v) \in \mathbb{R}^{C \times H' \times W'}$ from the mask M_v, where C, H' and W' are the number of channels, height and width. To capture the cross-view correlations [19], we introduce an Epipolar Cross Attention (ECA) module, which projects the anchor feature F_0 into the Query $Q \in \mathbb{R}^{C \times H' \times W'}$ and projects the auxiliary feature F_i into the Key and Value $K, V \in \mathbb{R}^{C \times H' \times W'}$. Then we define an epipolar vertical window $\mathcal{W}$ to address the probe bias, namely, each pixel p on the anchor view only needs to search within the corresponding window of the auxiliary view as:

$$\begin{aligned} \hat{F}(p) &= F_0(p) + \sum_{j \in \mathcal{W}(p)} \text{Softmax}\left(\frac{Q(p) \cdot K(j)}{\tau}\right) V(j), \\ \mathbf{z} &= \text{MLP}\left(\text{AvgPool}\left(\Phi_2(\hat{F})\right)\right), \end{aligned} \tag{2}$$

where τ is a temperature factor. $\hat{F}$ encodes the multi-view context. Φ_2 denotes the convolution. AvgPool$(\cdot)$ and MLP$(\cdot)$ denote average pooling and MLP.

2.2 Self-supervised SDF-tailored Differentiable Rendering

During test-time optimization, the model encounters clinical 2D echo masks that exhibit inherent domain gaps from the SSM priors. To enable patient-specific shape adaptation, we introduce a SDF-tailored differentiable rendering strategy. To address probe misalignment, we parameterize the transducer for each view v via a learnable spatial transformation $T_v = \{R_v, \mathbf{t}_v, \mathbf{s}_v\}$, denoting *rotation*, *translation*, and *scaling*, respectively. This allows us to explicitly map any sampled 2D pixel $p = (x, y)$ on the image plane ($\bar{p} = [x, y, 0]^\top$) to its corresponding 3D spatial coordinate $\mathbf{x}_p = \mathbf{s}_v R_v \bar{p} + \mathbf{t}_v$.

After querying the 3D SDF value at $\mathbf{x}_p$, the continuous distance should be converted into a 2D mask probability to compute the self-supervised loss. Traditional methods for surface extraction (*e.g.*, the step function) is non-differentiable and severs the gradient flow [10]. To resolve this, we formulate a differentiable rendering strategy that maps the predicted SDF value $f^{\theta}_{SDF}(\mathbf{z}, \mathbf{x}_p)$ to a soft

probability $\hat{M}_v(p)$ using a scaled Sigmoid function with a temperature parameter $\alpha > 0$ as:

$$\hat{M}_v(p) = \frac{1}{1 + \exp(\alpha \cdot f^{\theta}_{SDF}(\mathbf{z}, \mathbf{x}_p))}. \tag{3}$$

2.3 Cardiac Spatial-Temporal Reconstruction via Neural Velocity Modeling

We define a continuous neural velocity field Ψ_ω parameterized by a MLP. Given a 3D spatial point $\mathbf{x}_t \in \mathbb{R}^3$ and a normalized timestamp $t \in [0, 1]$, Ψ_ω predicts the velocity vector $\mathbf{v}_t \in \mathbb{R}^3$. Here, a sinusoidal positional encoding $\gamma(\cdot)$ [13] is applied to preserve high-frequency motion details:

$$\mathbf{v}_t = \Psi_\omega(\gamma(\mathbf{x}), \gamma(t)). \tag{4}$$

Furthermore, we employ a bidirectional LSTM [3] Φ_{BiL} to refine the temporal motion correlations of the latent vectors $[\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_T] = \Phi_{BiL}([\mathbf{z}_1, \ldots, \mathbf{z}_T]; \theta_{BiL})$. Therefore, we obtained the C-SDF f^{θ}_{SDF} for 3D shape and the velocity field Ψ_ω for cardiac motion. To overcome the inherent limitation of the SDF (topologically inconsistent meshes generated by frame-by-frame Marching Cubes [12]), we extract only a single initial 3D mesh $\mathbf{M}_0 = (\mathcal{X}_0, \mathcal{F})$ at $t = 0$ from f^{θ}_{SDF}, where $\mathcal{X}_0$ and $\mathcal{F}$ denote the initial vertex set and face topology. For subsequent frames, we propagate each individual vertex $\mathbf{x}_{t-1} \in \mathcal{X}_{t-1}$ using the Ψ_ω as $\hat{\mathbf{x}}_t = \mathbf{x}_{t-1} + \mathbf{v}_{t-1}$. To correct shape deviations accumulated by Ψ_ω, we introduce an explicit Radial SDF Alignment strategy leveraging the shape prior from each frame's SDF:

$$\mathbf{x}_t = \hat{\mathbf{x}}_t - f^{\theta}_{SDF}(\tilde{\mathbf{z}}_t, \hat{\mathbf{x}}_t) \cdot \frac{\nabla f^{\theta}_{SDF}(\tilde{\mathbf{z}}_t, \hat{\mathbf{x}}_t)}{\|\nabla f^{\theta}_{SDF}(\tilde{\mathbf{z}}_t, \hat{\mathbf{x}}_t)\|_2} \tag{5}$$

where $\hat{\mathbf{x}}_t$ and $\mathbf{x}_t$ denote the drifted vertices and the aligned vertices at frame t. The scalar $f^{\theta}_{SDF}(\tilde{\mathbf{z}}_t, \hat{\mathbf{x}}_t)$ evaluates the signed distance from the drifted vertex to the true surface, while $\frac{\nabla f^{\theta}_{SDF}}{\|\nabla f^{\theta}_{SDF}\|_2}$ represents the unit surface normal. By moving the drifted vertex $\hat{\mathbf{x}}_t$ along the inverse normal direction by the SDF distance, this operation explicitly pushes it onto the zero-level set ($f^{\theta}_{SDF} = 0$).

2.4 Model Optimization and Loss Function

We optimize our proposed Echo4DIR using two strategies: 1) SSM priors Supervised pretraining and 2) Test-time self-supervised learning.

SSM Priors Supervised Pretraining. We pretrain the CCSDF using SSM data [9,10], simulating US imaging to synthesize pseudo-echo masks by rotating the meshes to standard apical views. The CCSDF predicts the implicit surface at sampled 3D query points $\mathbf{x} \in \Omega$. The network is optimized end-to-end using an L_1 reconstruction loss and an L_2 regularization loss to prevent overfitting as:

$$\mathcal{L}_{pt} = \frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} \left| f^{\theta}_{SDF}(\mathbf{z}, \mathbf{x}) - s_{gt}(\mathbf{x}) \right|_1 + \lambda_1 |\mathbf{z}|_2^2, \tag{6}$$

where s_{gt} is the ground-truth SDF from SSM mesh and λ_1 is the weight scalar.

Test-time self-supervised learning. 1) *Shape* self-supervised learning. The cardiac shape is optimized by a hybrid loss $\mathcal{L}_{render}$, which combines pixel-wise Binary Cross-Entropy (BCE) $\mathcal{L}_{BCE}$ and global Dice Loss $\mathcal{L}_{Dice}$ between the rendered probability $\hat{M}_v$ and the clinical mask M_v:

$$\mathcal{L}_{render} = \frac{1}{V}\sum_{v=1}^{V}\Big(\frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}\mathcal{L}_{BCE}(\hat{M}_v(p), M_v(p)) + \mathcal{L}_{Dice}(\hat{M}_v, M_v)\Big). \quad (7)$$

2) *Motion* self-supervised learning. We supervise the velocity network Ψ_ω through the temporal evolving implicit surfaces. Based on Lagrangian kinematics [14,15], a physical point $\mathbf{x}_t$ advected by $\mathbf{v}_t$ should lie on the zero-level set of the next frame's surface. To ensure robust motion tracking, we introduce a bidirectional transport loss [14] as:

$$\begin{aligned}\mathcal{L}_{transport} =& \mathbb{E}_{\mathbf{x}_t}\Big[w(\mathbf{x}_t)\cdot|f^{\theta}_{SDF}(\tilde{\mathbf{z}}_{t+1}, \mathbf{x}_t + \mathbf{v}_t)|_2^2\Big] + \\ & \mathbb{E}_{\mathbf{x}_{t+1}}\Big[w(\mathbf{x}_{t+1})\cdot|f^{\theta}_{SDF}(\tilde{\mathbf{z}}_t, \mathbf{x}_{t+1} - \mathbf{v}_{t+1})|_2^2\Big] + \mathcal{L}_{smooth},\end{aligned} \quad (8)$$

where the weight $w(\mathbf{x}) = \exp(-|f^{\theta}_{SDF}(\tilde{\mathbf{z}}, \mathbf{x})|)$ ensures the optimization focuses on the anatomical boundaries. $\mathbb{E}_{\mathbf{x}_t}$ denotes the expectation over the sampled points $\mathbf{x}_t$. The smoothness loss $\mathcal{L}_{smooth} = \mathbb{E}_{\mathbf{x}\sim\mathcal{U}(\Omega)}\Big[\|\mathbf{v}_t(\mathbf{x})\|_2^2\Big]$ penalizes the squared velocity of uniformly sampled points $\mathcal{U}(\Omega)$.

3 Experiments and Results

3.1 Dataset and Implementation

Our framework is evaluated on two distinct datasets: 1) **SSM Dataset.** Following [9,10], we generated 10,000 3D left ventricular (LV) meshes (9,000 for training, 1,000 for testing) using a 18-mode SSM. To compute the SDF ground truth, 100,000 spatial points were sampled per mesh. 2) **Control Cohort Dataset.** We utilized a public clinical Control Cohort [10] comprising 144 normal transthoracic echocardiography subjects. Raw sequences were filtered for quality, temporally aligned to end-diastole, resized to 224×224, and the LV blood pool in apical views (A2C, A3C, A4C) was segmented via nnU-Net [8].

Evaluation Metrics. Our method is compared with the baseline approach MIA'25 [10], the SOTA GCN-based method for LV mesh reconstruction. We evaluate geometric precision on SSM dataset using four standard 3D metrics: *Mean Absolute Error* (MAE), *Hausdorff Distance* (HD), *Chamfer Distance* (CD), and *Root Mean Squared Error* (RMSE), scaled to physical dimensions (mm). For the CC dataset, we calculate the *Dice Coefficient* (Dice) and *Intersection over Union* (IoU) between the 2D projected masks and the clinical segmentation inputs.

Implementation Details. Our framework is implemented in PyTorch on an RTX 4090 GPU. Φ_1 and Φ_2 employ ResNet-18 with dimensions of (64, 128)

Table 1: Quantitative comparison of reconstruction performance, evaluating 3D geometric on SSM and 2D projection overlap on Control Cohort (CC) dataset.

Method	SSM Dataset				CC Dataset	
	MAE(mm)↓	RMSE (mm)↓	HD (mm)↓	CD (mm)↓	Dice (%)↑	IoU (%)↑
E-PiVox [18]	3.41 ± 0.85	3.75 ± 1.05	13.04 ± 4.52	7.34 ± 1.88	87.35 ± 3.48	77.97 ± 4.02
MIA'25 [10]	2.61 ± 0.78	2.86 ± 0.86	7.98 ± 1.43	5.61 ± 0.89	90.51 ± 2.72	84.76 ± 3.86
Echo4DIR	1.15 ± 0.13	1.33 ± 0.16	5.49 ± 0.54	2.84 ± 0.85	98.35 ± 0.42	96.75 ± 0.80

Table 2: Quantitative Ablation. $M_x \rightarrow M_y$ denotes the reconstruction of unseen view M_y from input view M_x.

Effectiveness of ECA and DR		
Method	MAE (mm) ↓	RMSE (mm) ↓
Echo4DIR	1.15 ± 0.13	1.33 ± 0.16
w/o ECA	1.28 ± 0.18	1.46 ± 0.21
w/o DR	1.54 ± 0.48	1.74 ± 0.59
Cross-View Geometric Reliability		
Views (ch)	Dice (%) ↑	IoU (%) ↑
4, 3 → 2	95.24 ± 1.40	90.94 ± 2.52
w/o DR	87.40 ± 5.37	81.87 ± 7.54
2 → 4, 3	94.89 ± 1.01	90.31 ± 1.81
w/o DR	84.73 ± 3.28	75.79 ± 4.33

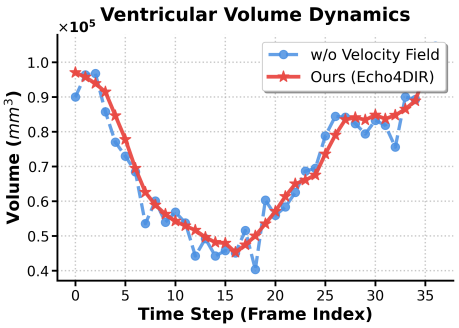

Fig. 2: Comparison of volume dynamics with and without the velocity field.

and (256, 512). The details of SDF follows [16]. Ψ_ω employ a 4-layer MLP with positional encoding. During 3D+t TTO, we perform 1,500 steps to optimize the first frame as the anchor, using an alternating shape-probe optimization (8:2 time ratio) to prevent geometric distortion from poor initial poses. Probe parameters are frozen for the next 2,000 steps of shape-motion optimization (random frame selection). We use the Adam optimizer with learning rates of 5×10^{-3}, 5×10^{-4}, and 1×10^{-3} for the probe, latent shape, and velocity fields. $\lambda_{BCE} = 1.0$, $\lambda_{Dice} = 0.5$, $\lambda_{trans} = 0.5$, and $\lambda_{smooth} = 0.001$. For rendering, 4,096 points are sampled per iteration, with the α increasing from 20 to 150. 3D and 4D reconstructions are completed within $\sim$ 30 s and 2 min.

3.2 Quantitative and Qualitative Results

Table 1 compares Echo4DIR with state-of-the-arts, including voxel-based E-PiVox [18] and mesh-based MIA'25 [10]. On the SSM dataset, Echo4DIR significantly outperforms MIA'25 (MAE: 2.61 →1.15 mm). To evaluate the clinical utility, we further validated on the CC dataset. Echo4DIR achieves superior performance with a 98.35% Dice. Such near-perfect overlap underscores the efficacy of our implicit representation in reconstructing geometric shape from clinical echo. In addition, Fig. 3 (b) and (c) visualize the 3D and 4D reconstructed meshes. Our method generates anatomically plausible cardiac surfaces

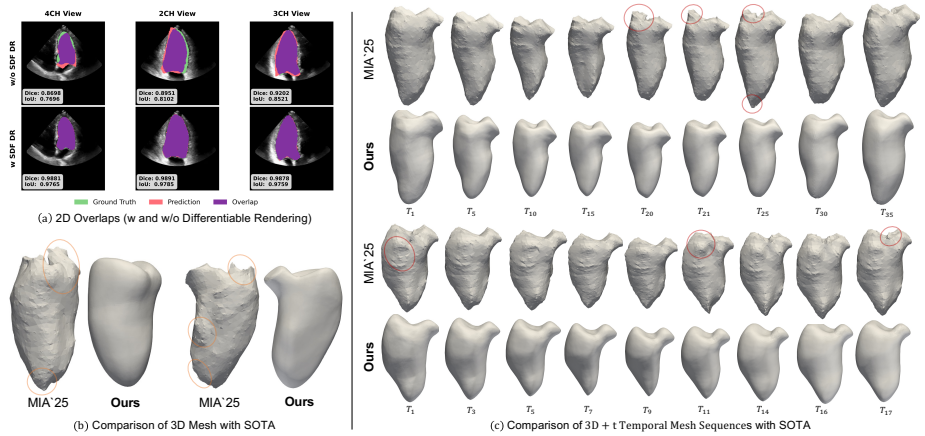

Fig. 3: Qualitative comparison of cardiac reconstruction performance.

with superior geometric smoothness, effectively eliminating unphysical artifacts and spikes. The 4D results capture consistent periodic cardiac motion.

Ablation. Table 2 validates the necessity of Eipolar Cross-Attention (ECA) and Differentiable Rendering (DR). Omitting ECA increases RMSE from 1.33 mm to 1.46 mm, proving the necessity of cross-view correlations. Removing DR causes RMSE to surge to 1.74 mm, demonstrating that gradient-based feedback is essential for robust shape adaptation. Fig. 2 illustrates the intuitive overlap alignment, demonstrating DR's effectiveness in processing real clinical data. As shown in Fig. 3, the velocity field effectively smooths the volume curve, capturing the flat isovolumic relaxation despite potential input segmentation errors.

Geometric Reliability. We further evaluate the reliability of our implicit prior by reconstructing unseen views from sparse inputs in Table 2. Notably, our framework achieves a remarkable 95.24% Dice for these unseen views, demonstrating its superior reconstruction fidelity. In contrast, performance without DR collapses significantly, underscoring its critical role in ensuring cross-view consistency and reliable shape reconstruction.

4 Conclusion

This work presents Echo4DIR, a 4D implicit framework that bridges the gap between sparse 2D clinical observations and high-fidelity cardiac anatomical geometry. We integrate conditional SDFs with a test-time differentiable rendering strategy to achieve patient-specific shape adaptation, yielding a remarkable 98.35% clinical Dice and 95.24% Dice in unseen view reconstruction, which validates the superior geometric reliability. The introduction of Radial SDF Alignment further ensures topological consistency during 3D+t extension, further eliminating non-physical artifacts. Future work will extend this framework to whole-heart modeling and leverage large-scale CT or MRI priors for pretraining.

In addition, we will further explore the utility and potential of these cardiac meshes for future digital twin integration such as surgical planning.

Acknowledgments. This work was supported by Singapore National Medical Research Council Open Fund - Young Individual Research Grant (25-1321-A0001), and Singapore Ministry of Education Tier 1 grant (25-1097-P0001). Y. Liu was partially supported by the China Scholarship Council (Grant No. 202407030035), National Natural Science Foundation of China (Grant No.62162068, 62462066, 62466060), Yunnan Province Ten Thousand Talents Program and Yunling Scholars Special Project (Grant No. YNWR-YLXZ-2018-022), Joint Fund Project for "Double First-Class" Construction of Science and Technology Department of Yunnan Province and Yunnan University (Grant No. 202301BF070001-025), Scientific Research Fund of Yunnan Provincial Department of Education (No.2026Y0185, 2025J0008), supported by No. FWCY-BSPY2024009, No. KC-252513122.

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

References

1. Boyd, D.P., Lipton, M.J.: Cardiac computed tomography. Proceedings of the IEEE **71**(3), 298–307 (2005)
2. Gao, H., Choi, H.F., Claus, P., Boonen, S., Jaecques, S., Van Lenthe, G.H., Van der Perre, G., Lauriks, W., D'hooge, J.: A fast convolution-based methodology to simulate 2-dd/3-d cardiac ultrasound images. IEEE transactions on ultrasonics, ferroelectrics, and frequency control **56**(2), 404–409 (2009)
3. Graves, A., Jaitly, N., Mohamed, A.r.: Hybrid speech recognition with deep bidirectional lstm. In: 2013 IEEE workshop on automatic speech recognition and understanding. pp. 273–278. IEEE (2013)
4. Haykin, S.: Neural networks: a comprehensive foundation. Prentice hall PTR (1994)
5. 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)
6. Hu, N., Yost, H.J., Clark, E.B.: Cardiac morphology and blood pressure in the adult zebrafish. The Anatomical Record: An Official Publication of the American Association of Anatomists **264**(1), 1–12 (2001)
7. Iacobellis, G., Ribaudo, M.C., Leto, G., Zappaterreno, A., Vecci, E., Di Mario, U., Leonetti, F.: Influence of excess fat on cardiac morphology and function: study in uncomplicated obesity. Obesity research **10**(8), 767–773 (2002)
8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
9. Laumer, F., Amrani, M., Manduchi, L., Beuret, A., Rubi, L., Dubatovka, A., Matter, C.M., Buhmann, J.M.: Weakly supervised inference of personalized heart meshes based on echocardiography videos. Medical image analysis **83**, 102653 (2023)
10. Laumer, F., Rubi, L., Matter, M.A., Buoso, S., Fringeli, G., Mach, F., Ruschitzka, F., Buhmann, J.M., Matter, C.M.: 2d echocardiography video to 3d heart shape reconstruction for clinical application. Medical Image Analysis **101**, 103434 (2025)

11. Li, L., Camps, J., Jenny Wang, Z., Beetz, M., Banerjee, A., Rodriguez, B., Grau, V.: Toward enabling cardiac digital twins of myocardial infarction using deep computational models for inverse inference. IEEE Transactions on Medical Imaging **43**(7), 2466–2478 (2024). https://doi.org/10.1109/TMI.2024.3367409
12. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. In: Seminal graphics: pioneering efforts that shaped the field, pp. 347–353 (1998)
13. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
14. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Occupancy flow: 4d reconstruction by learning particle dynamics. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5379–5389 (2019)
15. Osher, S., Sethian, J.A.: Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations. Journal of computational physics **79**(1), 12–49 (1988)
16. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 165–174 (2019)
17. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
18. Stojanovski, D., Hermida, U., Muffoletto, M., Lamata, P., Beqiri, A., Gomez, A.: Efficient pix2vox++ for 3d cardiac reconstruction from 2d echo views. In: International Workshop on Advances in Simplifying Medical Ultrasound. pp. 86–95. Springer (2022)
19. Wödlinger, M., Kotera, J., Keglevic, M., Xu, J., Sablatnig, R.: Ecsic: Epipolar cross attention for stereo image compression. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3436–3445 (2024)
20. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
21. Yu, J., Duan, Y., Huang, Y., Wang, Y., Ling, R., Luo, W., Zhang, A., Xu, J., Ni, Q., Zhou, Y., et al.: Ultratwin: towards cardiac anatomical twin generation from multi-view 2d ultrasound. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 608–617. Springer (2025)
22. Yuan, X., Liu, C., Wang, Y.: 4d myocardium reconstruction with decoupled motion and shape model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21252–21262 (2023)
23. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)