4D Myocardium Reconstruction with Decoupled Motion and Shape Model

Xiaohan Yuan Cong Liu Yangang Wang

Southeast University, China

1. More implementation details

Network architecture. Our network architecture is shown in Fig. 1.

The motion model refers to the multi-layer perceptron architecture [1, 2] to predict the deformation field of a given object from the space at arbitrary phase T_i to the ED spatial space. Specifically, for a sampling point $\mathbf{x} \in \mathbb{R}^3$, the corresponding deformation $\mathbf{v} \in \mathbb{R}^3$ is predicted under the condition of motion code $\mathbf{c}_i^{\mathrm{m}} \in \mathbb{R}^{K^{\mathrm{m}}} (K^{\mathrm{m}} = 128)$ at T_i phase and phase indicator $\tau_i \in \mathbb{R}$. The network has several jump connections, connecting the front features with the back. Except for the last layer, each layer has non-linear activation of LeakyReLU.

The shape model uses the auto-decoder network proposed in DeepSDF [3]. Precisely, for a point $\mathbf{x} \in \mathbb{R}^3$ in the space, the corresponding SDF value $sdf \in \mathbb{R}$ can be predicted through an MLP network based on the shape code $\mathbf{c}^{\mathrm{s}} \in \mathbb{R}^{K^{\mathrm{s}}} (K^{\mathrm{s}} = 256)$. In order to accelerate the convergence of the whole architecture, our shape model is pretrained.

Sampling strategy. For the training stage, we use the shape sequence obtained by the following method (in Sec. 3) for training and sample 50,000 points at each phase. The sampling points are mainly near the mesh surface. For the inference stage, the input is the point cloud sequence obtained from the raw CMR sequence. The point cloud is relatively dense at the intra-layer level, and each phase only contains about 5,000 contour points.

2. Acquisition of $\mathcal{P}$

Before entering the motion model, we must align the input point cloud into a unified space. The registration principle we adopt here can be seen in Fig. 2:

1. Align the normal vector of the short-axis slice of MR at the ED phase (red arrow) with that of the statistical mean shape $\bar{\mathbf{s}}$ (defined in Sec. 3.2.1 in main paper) to get $\mathcal{R}_1$.

2. Align the vertical line of the intersection between the left ventricle and the right ventricle of the middle slice of MR (green arrow) at the ED phase with that of the $\bar{\mathbf{s}}$ to get $\mathcal{R}_2$.

3. $\mathcal{R} = \mathcal{R}_2\mathcal{R}_1$ is computed as the result of the matrix multiplication. It forms the rotation part of $\mathcal{P}$.

Both $\mathcal{R}_1$ and $\mathcal{R}_2$ are computed using the following processing. Given two unit vectors, $\mathbf{u}_1$ and $\mathbf{u}_2$, the rotation matrix $\mathcal{R}_i$ that aligns $\mathbf{u}_1$ with $\mathbf{u}_2$ can be calculated as follows:

$$\mathcal{R}_i = \mathcal{I}_3 + \mathcal{V}_\times + \mathcal{V}_\times\mathcal{V}_\times \frac{1 - \mathbf{u}_1 \cdot \mathbf{u}_2}{\|\mathbf{u}_1 \times \mathbf{u}_2\|^2}, \tag{1}$$

where, $\mathcal{I}_3$ is the 3×3 identity matrix and $\mathcal{V}$ is the skew-symmetric cross-product matrix of $\mathbf{u}_1 \times \mathbf{u}_2$.

The obtained transformation $\mathcal{P}$ is executed together for the same sequence, normalized to the unit space, and then the points in this space are queried.

3. Our proposed dataset

We train and evaluate our proposed CMR dataset, which was acquired at Jiangsu Province Hospital, and composed of 55 healthy subjects. Each subject includes multiple slices (8-10 slices) with a resolution of $1.25 \times 1.25 \times 10mm$. Each slice covers the video sequence of the cardiac cycle (25 phases). The clinical experts manually delineated the left myocardium of all the phases and slices.

The acquisition of ground truth shape. First, the raw images are manually segmented to obtain the left myocardial contours $\mathbb{C} = \{\mathbf{C}_1, \mathbf{C}_2, \cdots, \mathbf{C}_M\}$, where M is the number of slices. We use the constructed statistical shape model (SSM) to fit $\mathbb{C}$. The formula for obtaining the shape from the given parameter $\theta \in \mathbb{R}^K (K = 100)$ is as follows:

$$\mathcal{B}(\theta, \mathbf{p}) = \mathbf{A}\mathbf{R}\mathcal{B}_s(\theta) + \mathbf{T}, \tag{2}$$

where the function $\mathcal{B}_s(\theta) = \mathcal{M}\bar{\mathbf{s}} + \mathbf{X}\theta$ produces the shapes according to the Principal Component Analysis (PCA). Noted that the hearts of different people are in different positions in space, so it is necessary to perform a global position transformation. Therefore, we add the position parameter $\mathbf{p}$ and limit the changes to scaling, rotation, and translation.

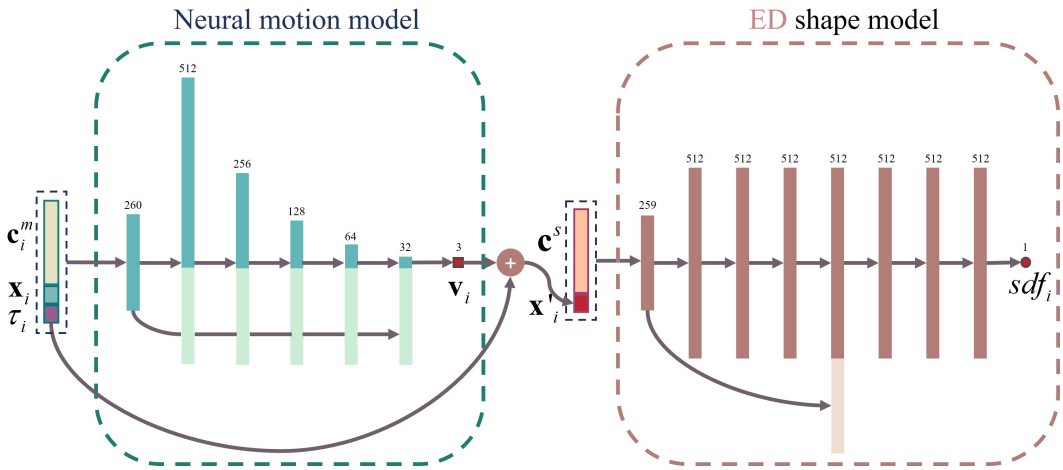

Figure 1. Network architecture.

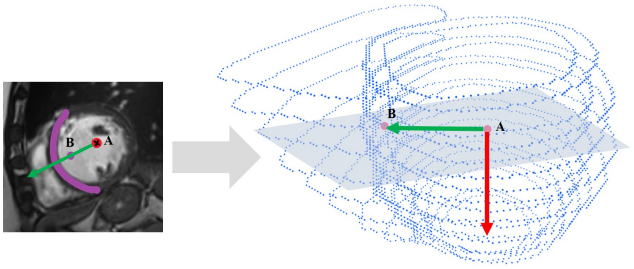

Figure 2. Align the input point cloud into unified space before entering the motion model.

$\mathbf{A}$ is a scaling matrix with the main diagonal elements a. $\omega = [\omega_1, \omega_2, \omega_3]^T$ is the axis-angle representation of the rotation. Using the Rodrigues formula, the axis-angle can be converted to a rotation matrix:

$$\mathbf{R} = exp[\omega] = \mathbf{I} + sin\,\|\omega\|\,[\hat{\omega}] + (1 - cos\,\|\omega\|)[\hat{\omega}]^2, \quad (3)$$

where $\hat{(\cdot)}$ is an operator that unites a vector, $[\cdot]$ is an anti-symmetric operator, and $\mathbf{I}$ is the identity matrix. $\mathbf{t} = [t_1, t_2, t_3]^T$ denotes the translation and $\mathbf{T}$ represents the matrix form used for calculation. Therefore, we define $\mathbf{p} = [a, a, a, \omega_1, \omega_2, \omega_3, t_1, t_2, t_3]^T \in \mathbb{R}^7$ as the position parameter.

We fit the template $\bar{\mathbf{s}}$ to the contour by formulating an optimization problem. In order to obtain the initial position better, we consider the relative position constraint between the left ventricle and the right ventricle. The 3D model obtained from the parameter is then sectioned according to the positions of the slices to find the correspondence between the points on the same slice. Therefore, our goal is to find the shape and position parameters by minimizing the square error between the contour of the reconstruction model $\widetilde{\mathbf{C}} = \prod \mathcal{B}(\theta, \mathbf{p})$ and $\mathbf{C}$, where $\prod$ is the section operation:

$$\begin{aligned} E = &\sum_{i=1}^{T_N}\sum_{j=1}^{M^{epi}}\sum_{k=1}^{N^{epi}} \|\widetilde{\mathbf{C}_{i,j,k}^{epi}} - \mathbf{C}_{i,j,k}^{epi}\|_2^2 \\ &+ \sum_{i=1}^{T_N}\sum_{j=1}^{M^{endo}}\sum_{k=1}^{N^{endo}} \|\widetilde{\mathbf{C}_{i,j,k}^{endo}} - \mathbf{C}_{i,j,k}^{endo}\|_2^2 \\ &+ \lambda \|\mathbf{N}\boldsymbol{\Theta}\|_2^2, \end{aligned} \quad (4)$$

where $\mathbf{C}_{i,j,k} \in \mathbb{R}^3$ and $\widetilde{\mathbf{C}_{i,j,k}} \in \mathbb{R}^3$ are the corresponding k-th vertice of $\mathbf{C}$ and $\widetilde{\mathbf{C}}$ on j-th slice at i-th phase respectively, which can be fastly searched by KD tree. epi and $endo$ represent epicardium and endocardium respectively, $\mathbf{L}$ is the Laplace matrix and $\boldsymbol{\Theta} = [\theta_1, \theta_2, \cdots, \theta_{T_N}] \in \mathbb{R}^{T_N \times (K+7)}$ is a matrix of all parameters of a sequence, and T_N is the length of the sequence. The last item is used for temporal smoothing. The whole process requires multiple iterations, and one sequence takes 30 minutes, which is far longer than the time required for practical application.

References

[1] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5932–5941, 2019. 1

[2] Boyan Jiang, Xinlin Ren, Mingsong Dou, Xiangyang Xue, Yanwei Fu, and Yinda Zhang. Lord: Local 4d implicit representation for high-fidelity dynamic human modeling. In *ECCV*, 2022. 1

[3] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 1