4D CardioSynth: Synthesising Dynamic Virtual Heart Populations through Spatiotemporal Disentanglement

Haoran Dou[1,2,3], Jinghan Huang[1,2,3], Arezoo Zakeri[1,2,4], Zherui Zhou[1,2,5], Tingting Mu[1,2,3], Jinming Duan[1,2,4], and Alejandro F. Frangi[1,2,3,4,6,7]

[1] Centre for Computational Imaging and Modelling in Medicine (CIMIM), University of Manchester, Manchester, UK
[2] Christabel Pankhurst Institute, University of Manchester, Manchester, UK
[3] Department of Computer Science, University of Manchester, Manchester, UK
[4] Division of Informatics, Imaging & Data Sciences, University of Manchester, Manchester, UK
[5] Department of Electrical & Electronic Engineering, University of Manchester, Manchester, UK
[6] NIHR Manchester Biomedical Research Centre, Manchester Academic Health Sciences Centre, University of Manchester, Manchester, UK
[7] Medical Imaging Research Centre (MIRC), Department of Cardiovascular Sciences and Department of Electrical Engineering, KU Leuven, Leuven, Belgium
alejandro.frangi@manchester.ac.uk

Abstract. Dynamic virtual populations are critical for realistic in-silico cardiovascular trials, yet current approaches primarily generate static anatomies, limiting their clinical and computational value. In this study, we present 4D CardioSynth, a generative framework for constructing dynamic 3D virtual populations of cardiovascular structures that change over time (3D+t). To model the complex interplay between cardiac structure and motion, we develop a factorised variational approach that disentangles spatial and temporal information in latent space, enabling independent control over anatomical variations and motion patterns. We demonstrate 4D CardioSynth's performance using a diverse dataset of bi-ventricle shapes acquired from 6,500 patients across complete cardiac cycles. Our results illustrate the superiority of 4D CardioSynth over state-of-the-art methods with respect to anatomical specificity, diversity, and generalisability, as well as motion plausibility. This approach enables more accurate virtual trials for cardiovascular interventions.

Keywords: Generative Model · Virtual Population · In-silico Trials.

1 Introduction

In-silico trials have emerged as powerful tools that offer a cost-effective and ethical alternative to traditional clinical trials. Virtual populations are a fundamental component of in-silico trials, aiming to provide a more diverse range of anatomical and physiological variations than the restricted real cohorts in clinical trials. These virtual populations are typically represented parametrically as a set of anatomical structures sampled from generative models [5]. In the cardiovascular domain, the generation of realistic patient-like cardiac anatomies that capture both spatial and temporal dynamics is crucial for ensuring the validity and generalisability of in-silico trials. However, creating such virtual populations of dynamic anatomy remains challenging due to the complex interplay between cardiac structure and motion, as well as the potential missing time point across the sequence in the dataset.

Statistical shape models have served as the traditional baseline for generating virtual populations [6, 8], but they are limited in their capacity to model complex spatiotemporal variations of cardiac shape. With the advent of deep learning, generative models have emerged as promising alternatives for capturing anatomical variability more accurately. Recent approaches focused on static shape modelling, where Beetz *et al.* proposed a variational autoencoder (VAE) to learn left ventricular shape variations [1], and Dou *et al.* enhanced this framework by incorporating normalising flows in the latent space to improve the model flexibility [4]. Whilst these methods demonstrated improved shape modelling capabilities, their limitation to fixed time points (end-diastole, end-systole or both) restricted their utility in motion-related in-silico studies. To address this limitation, a few studies explored dynamic cardiac anatomy generation. Qiao *et al.* introduced a VAE-based framework to model voxel-based cardiac anatomy sequences, employing recurrent neural networks (RNNs) to capture temporal dynamics [12], and later advancing to transformer architectures for improved spatiotemporal modelling [11]. However, the entanglement of spatial and temporal features in the latent space may constrain the model's flexibility, potentially reducing the diversity of virtual populations. On the other hand, they require the full sequence data to be available for each training iteration so that the latent RNN can capture the temporal information.

We introduce a novel generative framework, namely 4D CardioSynth, for synthesising dynamic 3D+t virtual heart populations with disentangled spatial and temporal representations. Our approach employs a specialised VAE architecture that explicitly decomposes the latent space into independent spatial and temporal subspaces, each processed by dedicated decoders for shape and motion reconstruction. We incorporate a temporal recurrent module to capture complex temporal dependencies, ensuring physiologically plausible cardiac motion patterns. We evaluate our framework comprehensively, assessing anatomical fidelity, diversity and generalisability, as well as motion plausibility through established metrics. Our experimental results demonstrate that the proposed framework generates more realistic cardiac anatomies with superior motion characteristics over the current state-of-the-art method, CHeart [12]. The disentangled representation enables independent control over anatomical variations and motion patterns, facilitating more flexible virtual population generation for in-silico trials.

2 Methodology

Our framework aims to generate dynamic 3D cardiac anatomies by explicitly disentangling spatial and temporal representations through a specialised variational autoencoder (VAE) architecture. As illustrated in Fig. 1, the framework consists of four main components: an encoder that extracts latent representations from input cardiac meshes, a decomposed latent space that separates spatial and temporal information, a temporal recurrent module (TRM) to capture the motion features and dual decoders that reconstruct anatomical shapes and motion patterns, respectively.

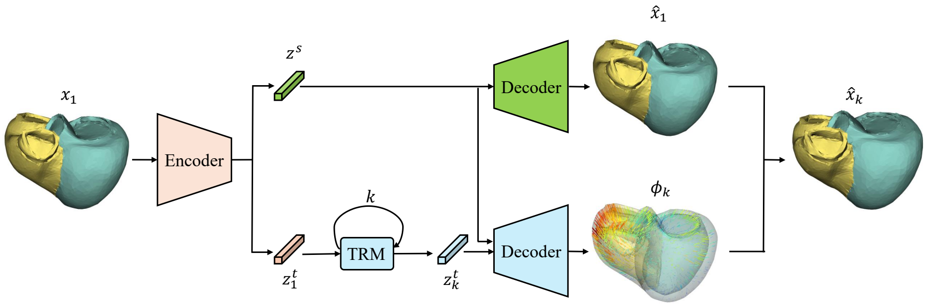

Fig. 1. Schematic illustration of our proposed 4D CardioSynth architecture.

2.1 Dynamic Virtual Heart Populations Synthesis

In our approach, a dynamic virtual heart sequence is represented as a sequence of triangular surface meshes. These meshes correspond to the cardiac structures of interest, i.e., Bi-Ventricle used in our study containing Left and Right Ventricles (BiV, LV and RV). Each mesh is defined by a set of 3D vertex coordinates along with an adjacency matrix that encodes vertex connectivity (i.e., the edges forming triangular faces). By performing template-based registration [14], we ensure that all meshes (regardless of the sample or time point) share a consistent topology and vertex ordering.

Within this framework, we denote the sequence of meshes as $X = \{x_k\}_{k=1}^{T}$, where x_k represents the 3D coordinates of all vertices at time point k. The proposed framework is designed to estimate the conditional distribution $p(x_k \mid x_1)$ of future heart shape x_k given the shape at end-diastole (ED) time point x_1. This is achieved by modeling the following joint distribution:

$$p(x_k, z_k^{\mathrm{t}}, z_1^{\mathrm{t}}, z^{\mathrm{s}}, x_1) = p_\theta(x_k \mid z_k^{\mathrm{t}}, z^{\mathrm{s}}, x_1) p_\gamma(z_k^{\mathrm{t}} \mid z_1^{\mathrm{t}}) q_\psi(z^{\mathrm{s}} \mid x_1) q_\psi(z_1^{\mathrm{t}} \mid x_1), \tag{1}$$

where the latent temporal and spatial codes, i.e., z_1^{t} and z^{s} computed from the static input by an encoder ψ as well as the temporal code z_k^{t} predicted at each time step by a neural network γ, are explained below.

Spatiotemporal disentanglement: We assume that a heart shape at a fixed time point (i.e., ED in this study) can be explicitly decomposed into two distinct latent components: a spatial code z^{s} and a temporal code z_1^{t}, modelled by the conditional distribution $q_\psi(z^{\mathrm{s}} \mid x_1^{\mathrm{t}})$ and $q_\psi(z_1^{\mathrm{t}} \mid x_1^{\mathrm{t}})$, respectively, parameterised by an encoder ψ. The spatial code z^{s} is intended to capture patient-specific anatomical features that remain constant throughout the cardiac cycle, while the temporal code z_1^{t} encodes the time-dependent motion state for ED.

To enforce the disentanglement, we decode z^{s} to reconstruct the input mesh at ED (producing $\hat{x}_1$) so that z^{s} retains detailed static shape information. In contrast, z_1^{t} is learned indirectly through the TRM rather than via direct decoding, which encourages z_1^{t} to encode only the dynamic aspects of cardiac motion.

Temporal recurrent module: The temporal code of the heart changes over time, which is modelled as a first-order Markov process, meaning each latent state depends only on the previous state. As a result, the k-step transition from z_1^{t} to z_k^{t} can be expressed as:

$$p_\gamma(z_k^{\mathrm{t}} \mid z_1^{\mathrm{t}}) = p(z_1^{\mathrm{t}}) \int \prod_{k'=1}^{k-1} p_\gamma(z_{k'+1}^{\mathrm{t}} \mid z_{k'}^{\mathrm{t}}) \, \mathrm{d}z_{1:k-1}^{\mathrm{t}}. \tag{2}$$

Starting from z_1^{t}, we iteratively apply the one-step transition $p_\gamma(z_{k'+1}^{\mathrm{t}} \mid z_{k'}^{\mathrm{t}})$ (parameterized by a multi-layer perceptron γ) to reach z_k^{t}, instead of directly predicting z_k^{t} in a single leap. This yields a smoother temporal transition in the latent space as compared to a direct multi-step prediction.

Vertex-wise motion prediction: We obtain the future mesh at different time step by deforming the current mesh x_1. Applying a deformation field (the displacement of each vertex) from time 1 to time k, denoted by ϕ_k, the predicted mesh coordinates is given by

$$x_k = x_1 \circ \phi_k. \tag{3}$$

As a result, instead of directly estimating the distribution of the absolute coordinates of a heart $p_\theta(x_k \mid z_k^{\mathrm{t}}, z^{\mathrm{s}})$, we learn the distribution of the deformation field $p_\theta(\phi_k \mid z_k^{\mathrm{t}}, z^{\mathrm{s}})$ conditioned on the latent codes, through learning a decoder θ. By focusing on vertex-wise deformations, 4D CardioSynth ensures that the predicted motion remains physiologically coherent, resulting in more realistic virtual cardiac anatomy sequences.

2.2 Network Architecture

Both the encoder ψ and decoder θ contain five residual graph convolutional blocks. Each block comprises two Chebyshev graph convolutions [3], each followed by instance normalisation and SiLU activation [7]. A residual connection is added between the input and output of each graph-convolutional block. The number of feature maps for each block is 16, 32, 32, and 64 in the encoder and inverted in the decoder. We employ hierarchical mesh down/up-sampling operations as proposed in CoMA [13]. The MLP γ for learning temporal representations contains two fully connected layers with hidden dimensions of 64 and SiLU activation.

2.3 Loss Functions

The encoder ψ for computing static latent codes, the multi-layer perceptron γ for predicting dynamic latent code, and the decoder θ for predicting the deformation are trained in a supervised fashion. An objective function modified from the original evidence lower bound in [9] is used. It consists of three components: reconstruction loss, motion loss, and Kullback–Leibler (KL) divergence loss. A reconstruction loss is used to measure the difference between the reconstructed mesh and the input ground-truth mesh:

$$\mathcal{L}_{\text{recon}} = \|x_1 - \hat{x}_1\|_1. \tag{4}$$

To encourage accurate estimation of motion pattern, we use L1-loss to penalise the difference between the predicted mesh and its ground truth:

$$\mathcal{L}_{\text{motion}} = \|x_k - \hat{x}_1 \circ \phi_k\|_1. \tag{5}$$

The KL loss is leveraged to measure the divergence between the approximate posterior and the prior distribution (i.e., standard Gaussian distribution):

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(q_\psi(z^{\text{s}}|x_0)\|p(z^{\text{s}})) + D_{\text{KL}}(q_\psi(z_0^{\text{t}}|x_0)\|p(z_0^{\text{t}})). \tag{6}$$

Finally, the total loss is a weighted sum of these components:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{recon}} + \lambda_2\mathcal{L}_{\text{motion}} + \lambda_3\mathcal{L}_{\text{KL}}. \tag{7}$$

The setting of $\lambda_1 = 1$, $\lambda_2 = 2e^{-3}$, and $\lambda_3 = 1$ for weighting coefficients is employed, balancing the contribution of each term.

3 Experiments and Results

3.1 Dataset

In this study, we created a cohort of 6500 triangular meshes of the bi-ventricle of the heart based on a subset of cardiac cine-MR imaging data available from the UK Biobank (UKBB). The patient-specific meshes were reconstructed using [15]. Each patient contains 50 time points that cover the full cardiac cycle. We randomly split the dataset into 5000/500/1000 for training, validation, and testing, respectively.

3.2 Implementation Details

The framework was implemented using PyTorch on a standard PC with an NVIDIA RTX 4090 GPU with 24GB memory. We trained our model using the AdamW optimiser with an initial learning rate of 5e-4 and batch size of 128 for 500 epochs. The spatial and temporal latent dimensions were set at 16. The down/up-sampling factor was four.

3.3 Evaluation Metrics

We compared 4D CardioSynth with the state-of-the-art method, CHeart [12], which uses recurrent neural networks to model joint spatiotemporal representations without disentanglement. We also compared our method to a vanilla VAE [9] trained exclusively on the same datasets with solely single time points (ED) to benchmark spatial representations without input from motion patterns. Three metrics were used in the evaluation: (1) specificity [2], measured by the distance between generated meshes and their nearest neighbours in the real population; (2) diversity [10], measured by the fraction of real samples whose neighbourhoods contain at least one virtual sample; and (3) generalisability [5], formulated as the reconstruction error between the reconstructed and unseen shapes. The Euclidean distance metric was used in all metric calculations.

3.4 Results

Quantitative Analysis of Spatial Features: Our quantitative analysis of the spatial features of virtual populations generated by investigated methods is summarised in Table 1, focusing on the ED time point to isolate spatial aspects of cardiac morphology. It can be observed that 4D CardioSynth demonstrated superior performance compared to CHeart across all metrics and structures. With a biventricular specificity of 2.27±0.49 mm compared to 3.12±1.61 mm of CHeart, our method achieved better fidelity, indicating that the shapes generated by our model more closely resemble the anatomical characteristics observed in real populations. This improvement was consistent across individual chambers. Regarding diversity, 4D CardioSynth achieved 48.3% for BiV compared to CHeart (41.3%), with a 7% improvement in the anatomical variability that models capture from the real population. Similar improvements were observed for individual ventricles as well. Such results suggest that our disentangled latent space approach allows more flexible modelling of diverse cardiac anatomical structures. The results of generalisability evaluation also showed improvements with our method, achieving 1.58±0.43 mm compared to 1.73±0.58 mm for CHeart. This consistent improvement across all three metrics highlights the advantage of our spatiotemporal disentanglement approach. Notably, the conventional VAE model, which was trained exclusively on ED time point data, demonstrated comparable performance in all metrics compared to our method, indicating that 4D CardioSynth can avoid the degradation typically caused by the spatiotemporal entanglement in the latent representation.

Temporal Prediction Performance: Further temporal analysis revealed consistent performance advantages of our method throughout the cardiac cycle. Figure 2 illustrates the predictive error across all future time points given the input of the ED time point for CHeart and 4D CardioSynth. The prediction referred to here was formulated as a sequence completion task where the model takes shape at the first time point (ED) as input and predicts the shapes of the full cardiac cycle. Our model consistently achieved lower predictive errors than the CHeart

Table 1. Quantitative analysis on the spatial features of virtual populations generated by the investigated methods. The specificity and generalisability error are illustrated as mean±std

Methods	Specificity (mm, ↓)			Diversity (%, ↑)			Generalisability (mm, ↓)		
	LV	RV	BiV	LV	RV	BiV	LV	RV	BiV
VAE [9]	1.88±0.45	2.08±0.40	2.17±0.45	49.7	43.7	45.2	1.58±0.53	1.82±0.54	1.70±0.50
CHeart [12]	2.84±1.61	3.01±1.66	3.12±1.61	43.5	40.4	41.3	1.56±0.62	1.88±0.59	1.73±0.58
4D CardioSynth	1.98±0.48	2.17±0.45	2.27±0.49	49.2	44.7	48.3	1.47±0.44	1.68±0.48	1.58±0.43

throughout the cardiac cycle. This performance gap is particularly pronounced during the early time points and late time points, corresponding to diastole phases where anatomical precision is clinically essential. Both methods exhibited increased predictive errors during the mid-cycle time points, which coincide with the rapid deformation phases during systole. However, our method maintained a more stable performance with a narrower error variance. This stability can be attributed to our spatiotemporal disentanglement strategy, which effectively disentangles the complex cardiac motion patterns from spatial anatomical features.

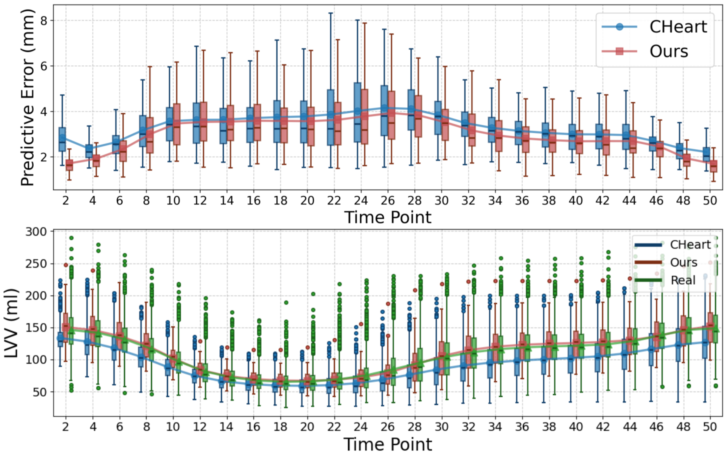

Fig. 2. Temporal analysis of the investigated methods. The top panel displays the distributional comparison of predictive error across the cardiac cycle between CHeart and our method. The bottom panel shows the LVV changes across the cardiac cycle of the real distribution and virtual populations generated by our method and CHeart.

Clinical Relevance of Virtual Populations: To assess the clinical relevance of the generated dynamic 3D+t virtual heart population, we measured each virtual patient's cardiac parameters (i.e., left ventricle volume, LVV). Figure 2 presents the LVV measurements across all 50 time points for the real cohort, virtual populations generated by 4D CardioSynth and CHeart. Both generative models captured the pattern of ventricular volume changes during the cardiac cycle, with volume reduction during systole and subsequent filling during diastole. During the diastole phase, virtual hearts generated by 4D CardioSynth more accurately represented the higher LVV and their associated variability across the population. The median LVV values from our synthetic population closely tracked the real data distribution, whereas the virtual population from CHeart slightly underestimated volumes during these phases. Similarly, during the systole phase, 4D CardioSynth better captured the reduced variability observed in the real data. The temporal coherence of volume changes in our virtual population also more closely resembled the pattern observed in real cardiac cycles, with smoother transitions between phases. This improved temporal modelling can be attributed to our spatiotemporal disentanglement approach, which effectively learns the underlying dynamics of cardiac motion for generating realistic virtual hearts.

Visual Assessment of Generated Virtual Hearts To demonstrate the quality of our generated virtual patient of the heart anatomy, Figure 3 visualises a representative heart shape generated by our method across different time points of the cardiac cycle. The visualisation shows the plausibility of the generated cardiac shape and the myocardium smoothly deforming through systole and diastole.

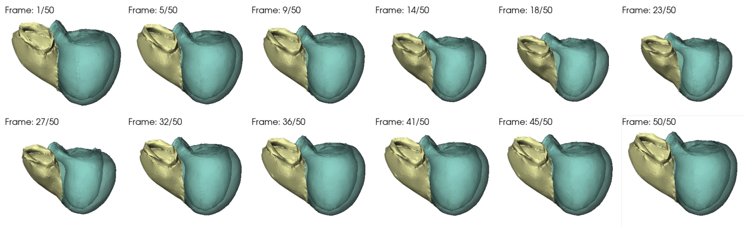

Fig. 3. Visualisation of a representative heart shape generated by our method at selected time points throughout the cardiac cycle.

4 Conclusion

In this study, we presented 4D CardioSynth for synthesising dynamic 3D+t virtual heart populations for in-silico trials. To learn the complex interplay between cardiac structure and motion, we introduced a disentangled latent space within a specialised VAE that independently captures spatial and temporal representations. This design enables separate control over anatomical variations and motion patterns, facilitating more flexible and targeted virtual population generation. We compared our method to CHeart, with results demonstrating that our approach achieves better fidelity, diversity, and generalisability in terms of spatial patterns, as well as the plausibility of the cardiac motion. Future work will focus on extending the current anatomical representation to encompass the full heart and more detailed structures, such as heart valves.

Acknowledgments. This research was carried out using data from the UK Biobank (access application 11350). This work was supported by the Royal Academy of Engineering, United Kingdom under the RAEng Chair in Emerging Technologies (INSILEX CiET1919/19), the ERC Advanced Grant – UKRI Frontier Research Guarantee (INSILICO EP/Y030494/1), the UK Centre of Excellence on in-silico Regulatory Science and Innovation (UK CEiRSI) (10139527), the National Institute for Health and Care Research (NIHR) Manchester Biomedical Research Centre (BRC) (NIHR203308), the BHF Manchester Centre of Research Excellence (RE/24/130017), and the CRUK RadNet Manchester (C1994/A28701).

Disclosure of Interests. AFF is a Founder, Board Member and Shareholder of adsilico ltd and OculomeX Health Ltd and consults with both companies. AFF has non-financial collaborative agreements with several medical device manufacturers. AFF and HD have patents licensed to adsilico or OculomeX and receive royalties from these arrangements. AFF also receives royalties from Elsevier associated with a medical image computing textbook. All other authors declare no competing interests.

References

1. Beetz, M., Corral Acero, J., Banerjee, A., Eitel, I., Zacur, E., Lange, T., Stiermaier, T., Evertz, R., Backhaus, S.J., Thiele, H., et al.: Interpretable cardiac anatomy modeling using variational mesh autoencoders. Frontiers in Cardiovascular Medicine **9**, 983868 (2022)
2. Davies, R.H., Twining, C.J., Cootes, T.F., Taylor, C.J.: Building 3-d statistical shape models by direct optimization. IEEE Transactions on Medical Imaging **29**(4), 961–981 (2009)
3. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. Advances in neural information processing systems **29** (2016)
4. Dou, H., Ravikumar, N., Frangi, A.F.: A conditional flow variational autoencoder for controllable synthesis of virtual populations of anatomy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 143–152. Springer (2023)
5. Dou, H., Virtanen, S., Ravikumar, N., Frangi, A.F.: A generative shape compositional framework to synthesize populations of virtual chimeras. IEEE Transactions on Neural Networks and Learning Systems (2024)
6. Frangi, A.F., Rueckert, D., Schnabel, J.A., Niessen, W.J.: Automatic construction of multiple-object three-dimensional statistical shape models: Application to cardiac modeling. IEEE transactions on medical imaging **21**(9), 1151–1166 (2002)
7. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)

8. Hoogendoorn, C., Sukno, F.M., Ordás, S., Frangi, A.F.: Bilinear models for spatio-temporal point distribution analysis: Application to extrapolation of left ventricular, biventricular and whole heart cardiac dynamics. International Journal of Computer Vision **85**(3), 237–252 (2009)
9. Kingma, D.P., Welling, M., et al.: Auto-encoding variational bayes (2013)
10. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: International conference on machine learning. pp. 7176–7185. PMLR (2020)
11. Qiao, M., McGurk, K.A., Wang, S., Matthews, P.M., Regan, D.P., Bai, W.: A personalised 3d+ t mesh generative model for unveiling normal heart dynamics. arXiv preprint arXiv:2409.13825 (2024)
12. Qiao, M., Wang, S., Qiu, H., De Marvao, A., O'Regan, D.P., Rueckert, D., Bai, W.: Cheart: A conditional spatio-temporal generative model for cardiac anatomy. IEEE transactions on medical imaging **43**(3), 1259–1269 (2023)
13. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3d faces using convolutional mesh autoencoders. In: Proceedings of the European conference on computer vision (ECCV). pp. 704–720 (2018)
14. Ravikumar, N., Gooya, A., Frangi, A.F., Taylor, Z.A.: Generalised coherent point drift for group-wise registration of multi-dimensional point sets. In: Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20. pp. 309–316. Springer (2017)
15. Xia, Y., Chen, X., Ravikumar, N., Kelly, C., Attar, R., Aung, N., Neubauer, S., Petersen, S.E., Frangi, A.F.: Automatic 3d+ t four-chamber cmr quantification of the uk biobank: integrating imaging and non-imaging data priors at scale. Medical Image Analysis **80**, 102498 (2022)