Mask2Surface: Motion Correction and Super-Resolution for Cardiac Surface Reconstruction Using Latent Diffusion

Zichen Zhang[1], Zhentao Liu[1], Zeng Zhang[1], and Zhiming Cui[1]

School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China
cuizhm@shanghaitech.edu.cn

Abstract. Cardiac magnetic resonance (CMR) imaging is one of the most important imaging modalities for cardiac analysis. However, short-axis CMR imaging can only produce a sparse set of 2D images with an extremely low inter-slice resolution. Moreover, these 2D slices are usually misaligned due to the respiratory and cardiac motion of the patients, strongly affecting the diagnosis and intervention procedures for cardiac diseases. Deep learning-based approaches have been proposed to tackle these problems, but they mostly focus on voxel representation, yielding rough cardiac surfaces that are difficult to analyze. Therefore, we propose a deep learning-based method to perform CMR motion correction and super-resolution simultaneously to acquire high-fidelity left ventricular myocardial surfaces. Given a set of 2D misaligned sparse segmentation masks of the left ventricular myocardium, our method first leverages an end-to-end convolutional neural network to correct and super-resolve the masks to approach the distribution of the motion-free and high-resolution masks. Then, the acquired super-resolved segmentation masks are estimated to form coarse signed distance grids, guiding a latent diffusion model to produce the corresponding high-fidelity myocardial surfaces. The superior performances of our approach are testified through comprehensive experiments in both simulation and clinical settings.

Keywords: Cardiac magnetic resonance · Motion correction · Super-resolution.

1 Introduction

Cardiac magnetic resonance (CMR) imaging is one of the gold standards for cardiac assessments. Clinically, short-axis CMR imaging is widely applied to acquire a sparse set of 2D cardiac image slices from multiple breath holds, bringing two unavoidable defects. One of the defects is the data sparsity, which means the inter-slice resolution of short-axis CMR slices is extremely low (only 8-10 mm). Another defect is the motion artifacts induced by the respiratory and cardiac motion of the patients during the breath holds, causing the already sparse

2D slices to be spatially misaligned. These defects strongly hinder the effective utilization of short-axis CMR imaging and hence need to be addressed.

In recent years, deep learning has been applied to CMR motion correction and super-resolution to achieve better results. Voxel-based methods [19,4,5] focus on segmentation masks from a generative [19] or discriminative [4,5] perspective. The generative method SRHeart [19] performs latent optimization based on a variational autoencoder (VAE) [7] to search in its latent space for proper latent vectors to generate the targeted high-resolution 3D cardiac masks. The discriminative approaches aim to build multi-stage [4] or end-to-end [5] pipelines for motion correction and super-resolution using the convolutional neural network (CNN). In addition, there is an approach [2] formulating CMR motion correction and super-resolution as a point cloud completion task, where the contours obtained from the segmentation masks are used to produce dense cardiac point clouds via the point completion network (PCN) [20]. However, their method has to rely on post-processing to generate meshes from the point clouds, leading to additional parameter tuning and constantly low-quality meshes.

The previous methods are all based on explicit shape representations, namely voxels, meshes, and point clouds. The voxel-based approaches can only produce rough surfaces that are difficult for analysis, and the method built upon point clouds has to be accompanied by additional mesh generation algorithms to form surfaces. Therefore, we set our sights on the implicit representation of the signed distance field (SDF) [11] to enable smooth surface reconstruction.

In this paper, we propose a deep learning-based method to perform motion correction and super-resolution jointly for CMR imaging to reconstruct high-fidelity left ventricular myocardial surfaces. In contrast to previous approaches, including DeepSDF [11], that utilize multi-layer perceptrons to model continuous SDFs, we represent cardiac surfaces in the format of signed distance grids (SDGs), which are SDFs discretely stored in voxel grids and enable the uniform and efficient utilization of CNNs for network architectures. These surface SDGs are compressed to a compact latent space by a vector-quantized variational autoencoder (VQ-VAE) [18] for a latent diffusion model [14]. The diffusion model is guided by coarse signed distance grids (C-SDGs) estimated from segmentation masks to learn the latent representations of the surface SDGs. To reconstruct surfaces from clinically acquired misaligned low-resolution segmentation masks, a CNN is trained to correct and super-resolve the masks, which are then utilized to estimate the C-SDGs. C-SDGs will guide the latent diffusion model to generate fine-grained SDGs, which will be converted into high-fidelity myocardial surfaces via Matching Cubes. The utilization of surface SDGs and C-SDGs as conditions is closely synergistic with the employed VQ-VAE encoder, maintaining promising reconstruction results even in challenging clinical settings. Comprehensive experiments have demonstrated that our method achieves better reconstruction quality both qualitatively and quantitatively.

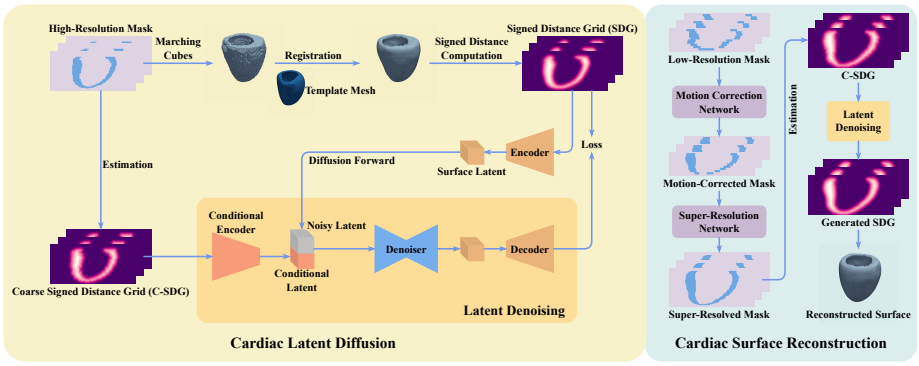

Fig. 1. An overview of our method. High-fidelity cardiac surfaces acquired by template mesh registration are converted to signed distance grids (SDGs) and compressed by the VQ-VAE to learn a compact latent representation. Coarse SDGs (C-SDGs) are estimated to serve as the conditions for the diffusion model to learn the surface latent. For surface reconstruction, the misaligned low-resolution masks are corrected and super-resolved by a CNN to produce the super-resolved masks, based on which the C-SDGs are estimated to guide the denoising process of the diffusion model, producing the denoised latent to reconstruct the surfaces.

2 Methods

Our method comprises two parts. The first part involves the training of a VQ-VAE and a latent diffusion model to learn the latent representations of high-fidelity cardiac surfaces, as shown in the left component of Fig. 1. The second part is illustrated in the right block of Fig. 1, depicting the cardiac surface reconstruction procedure. A CNN is trained to correct and super-resolve the clinically acquired misaligned low-resolution segmentation masks. Coarse signed distance grids are estimated from the masks to guide the trained diffusion model for conditional denoising, yielding surface latent to reconstruct the cardiac surfaces.

2.1 Cardiac Latent Diffusion

To reconstruct high-fidelity cardiac surfaces, we leverage a latent diffusion model to capture priors of high-quality surfaces based on the implicit representation of signed distance fields.

Cardiac Surface Compression. To ensure that the latent diffusion can learn a compact latent representation of high-fidelity cardiac surfaces, we first apply Marching Cubes [9] to the high-resolution segmentation masks to obtain the meshes. Next, a template cardiac mesh [1] is registered to these meshes, serving as the ground-truth surfaces. Then, the signed distances from the voxel points of the segmentation masks to the fitted cardiac meshes are computed to acquire the signed distance grids (SDGs), which are the discretely stored SDFs. We

train a 3D CNN-based VQ-VAE [18] to compress and reconstruct these grids, yielding a compact latent space of high-quality cardiac surfaces. The SDG and its reconstruction counterpart are denoted as $\mathbf{S}$ and $\hat{\mathbf{S}}$. The surface latent and the vector-quantized latent are denoted as $\mathbf{Z}^s$ and $\tilde{\mathbf{Z}}^s$. The loss function for VQ-VAE training is formulated as:

$$\mathcal{L}_{VQ} = \mathcal{L}_1(\hat{\mathbf{S}}, \mathbf{S}) + \left\| \text{sg}[\mathbf{Z}^s] - \tilde{\mathbf{Z}}^s \right\|_2^2 + \left\| \text{sg}[\tilde{\mathbf{Z}}^s] - \mathbf{Z}^s \right\|_2^2 \tag{1}$$

where $\text{sg}[\cdot]$ is the stop-gradient operation.

Latent Diffusion Training. After the VQ-VAE is trained, we freeze its weights and use its produced surface latent to train the latent diffusion model. For the conditioning purpose, we propose to utilize the coarse signed distance grids (C-SDGs) estimated from the segmentation masks. Given a segmentation mask $\mathbf{M}$, we first apply Marching Cubes [9] to obtain its corresponding mesh. Then, a C-SDG $\mathbf{C}$ is defined as a voxel grid with the same shape as $\mathbf{M}$. For each voxel position i of $\mathbf{C}$, the contained value is estimated via the following scheme:

$$\mathbf{C}[i] = \begin{cases} -1 \times \mathcal{D}(i) & \text{if } \mathbf{M}[i] = 1 \\ +1 \times \mathcal{D}(i) & \text{if } \mathbf{M}[i] = 0 \end{cases} \tag{2}$$

where $\mathbf{M}[\cdot]$ denotes querying the segmentation label at the voxel i of $\mathbf{M}$. The background voxels and myocardial voxels are labeled as 0 and 1, respectively. $\mathcal{D}(i)$ is the distance from the voxel i to the surface of the corresponding mesh. These estimated SDGs are described as C-SDGs because they contain inaccuracies in both signs and distance values. A conditional encoder will encode these C-SDGs into conditional latent $\tilde{\mathbf{Z}}^c$ for the diffusion model. Although C-SDGs are inherently inaccurate, they share similar semantics to the SDGs of the high-fidelity cardiac surfaces because every value within them represents the signed distances from voxel positions to the surfaces. Consequently, we directly employ the encoder of the VQ-VAE as the conditional encoder as it has been trained to encode the surface SDGs.

Given the vector-quantized and compressed latent representations of the high-fidelity cardiac surfaces $\tilde{\mathbf{Z}}^s$ and the corresponding conditional latent $\tilde{\mathbf{Z}}^c$, the latent diffusion model ϵ_θ is trained with the following loss function:

$$\mathcal{L}_{diff} = \mathbb{E}_{\tilde{\mathbf{Z}}^s, \tilde{\mathbf{Z}}^c, \epsilon, t} \left[\left\| \epsilon - \epsilon_\theta(\tilde{\mathbf{Z}}_t^s, \tilde{\mathbf{Z}}^c, t) \right\|_2^2 \right] \tag{3}$$

where $\tilde{\mathbf{Z}}_t^s$ is the noised latent representation. ϵ is a Gaussian noise variable and t is uniformly sampled from $\{1, ..., T\}$.

2.2 Cardiac Surface Reconstruction

The clinically acquired segmentation masks from CMR imaging are misaligned and have low resolutions. Therefore, a CNN is applied to correct and super-resolve these low-resolution masks. Then, we estimate the C-SDGs and reconstruct the surfaces based on a conditional denoising procedure.

Motion Correction and Super-Resolution Network. Following [5], we build and train a CNN-based network MCSR-Net with a motion correction network (MC-Net) and a super-resolution network (SR-Net) sequentially connected. It first produces translation vectors to realign the misaligned segmentation masks and then super-resolves them to increase the mask resolution. MC-Net follows the structure in the method [5] except that the ResBlock channels are reduced to 16, 32, 64, and 64. In contrast to the method [5] that processes full-resolution voxel grids in the super-resolution part, a modification is made by applying intra-slice downsampling and upsampling to increase the receptive fields and speed up computations. Besides, the channels are also decreased to constrain over-fitting. The loss function employed to train the MCSR-Net follows the method [5].

Surface Reconstruction by Conditional Denoising. Once the latent diffusion model and MCSR-Net are trained, cardiac surfaces can be reconstructed via a conditional denoising procedure. Given a misaligned low-resolution segmentation mask, we first feed it into the MCSR-Net to obtain the motion-free and super-resolved mask, based on which the C-SDG is estimated for the conditional encoder to produce the conditional latent. The conditional latent guides the diffusion model to iteratively denoise a Gaussian noise variable to obtain the desired surface latent, which is forwarded to the decoder of the VQ-VAE to generate the SDG. The cardiac surface can be extracted via Marching Cubes [9] at the zero iso-surface of the generated SDG.

3 Experiments and Results

3.1 Experimental Settings

Datasets. We utilize the cardiac super-resolution label maps dataset [15] for the experiments. It contains motion-free high-resolution segmentation masks from 3D balanced steady-state free precession cine sequences and motion-corrupted low-resolution masks obtained from multiple breath holds of 1331 patients. We divide them into 1024 for training, 72 for validation, and 235 for testing. The high-resolution masks are resampled to a voxel spacing of 1 mm $\times$ 1 mm $\times$ 2 mm and then centered and cropped to 128 $\times$ 128 $\times$ 60. Two experimental settings, namely simulation and clinical settings, are employed to conduct a comprehensive study. (1) For the simulation setting, the misaligned low-resolution masks are simulated by first 5$\times$ down-sampling the high-resolution masks in the slice dimension and then performing slice-wise translations using two-dimensional randomly generated vectors. Each entry of the vectors is generated independently using a Gaussian distribution with a mean of 3.45 mm and a standard deviation of 1.305 mm. This is the distribution fitted in a previous study [17] but multiplied by a factor of 1.5 to include higher diversity. For each high-resolution mask, 12 misaligned low-resolution masks are simulated. (2) For the clinical setting, we directly utilize the provided realistic low-resolution masks in the dataset.

Implementation Details. The mesh registration and signed distance computation are implemented using PyTorch3D [13] and Open3D [21]. For mesh registration, an MLP is employed to deform the template to fit the target meshes with the mesh vertex coordinates as inputs and the deformation vectors as outputs. The loss function consists of Chamfer distance, Laplacian smoothing, and L2 norm of the vectors. The VQ-VAE, with the codebook size 1024 and latent dimension 8, is trained for 800 epochs with an initial learning rate of 0.0001 and a reduction-on-plateau scheduler using a decay factor of 0.9 based on validation loss. The diffusion U-Net from MONAI Generative [3,12] is used as the denoiser with 128, 256, and 384 channels at each depth. It is trained for 3000 epochs using 1000-step DDPM [6] and 1e-5 as the base learning rate with the same scheduler as the VQ-VAE. The conditional encoder is frozen while training diffusion. For reconstruction, DDIM [16] with 10 sampling steps is used for denoising. DDPM and DDIM employ a linear schedule between 0.0015 and 0.0195. All the optimization processes are performed using Adam optimizer [8] on an NVIDIA 4090 GPU.

Competing Methods and Evaluation Metrics. Five methods are included as the competing methods, namely nearest-neighbor interpolation (NNI), SRHeart [19], MCSR-MS [4], MCSR-ETE [5], and CardiacPCN [2]. We utilize the Dice coefficient as the primary evaluation metric. Since CardiacPCN is a point cloud-based method producing non-watertight meshes constantly, Chamfer distance (CD) is leveraged as an additional metric.

3.2 Experimental Results

Quantitative Results. The methods are first trained and tested using the simulated low-resolution masks and the respective high-resolution masks. The experimental results are listed in Table 1. In this simulation setting, our method achieves the best shape accuracy among these approaches. For the clinical setting, two experiments are conducted using the clinical paired low-resolution and high-resolution masks provided in the dataset. We first directly test the methods trained on the simulated masks using the clinical masks. Next, we finetune these methods with the training set of clinical masks and then test them on the test set. For our method, only the MCSR-Net is finetuned. The quantitative results are listed in Table 2. In both experiments, our method reaches the best reconstruction accuracies. These methods usually experience a performance drop when moving from simulation to clinical settings, showing that there is still room to improve the generation and utilization of simulated data.

Qualitative Results. The visualized reconstructed surfaces are shown in Fig. 2. The surfaces produced by NNI, SRHeart [19] MCSR-MS [4], and MCSR-ETE [5] are rough as they are voxel-based methods, forming surfaces by Marching Cubes with an iso-value of 0.5. The surfaces from NNI preserves significant motion

Table 1. The performances of the methods in the simulation setting.

Methods	Dice (%)	CD (mm)
NNI	64.77 ± 3.91	21.94 ± 2.90
SRHeart [19]	76.89 ± 4.05	18.41 ± 4.61
MCSR-MS [4]	91.24 ± 3.09	6.74 ± 1.51
MCSR-ETE [5]	92.97 ± 2.95	6.30 ± 1.41
CardiacPCN [2]	-	13.16 ± 83.08
Ours	**94.05 ± 2.05**	**5.59 ± 1.13**

Table 2. The performances of the methods in the clinical settings.

Methods	Direct		Finetuning	
	Dice (%)	CD (mm)	Dice (%)	CD (mm)
NNI	65.84 ± 6.57	22.08 ± 5.34	65.84 ± 6.57	22.08 ± 5.34
SRHeart [19]	74.43 ± 6.61	19.85 ± 7.43	74.43 ± 6.61	19.85 ± 7.43
MCSR-MS [4]	73.69 ± 5.08	16.54 ± 3.94	84.19 ± 4.96	11.34 ± 5.21
MCSR-ETE [5]	74.23 ± 4.90	16.32 ± 4.11	84.21 ± 5.12	11.19 ± 5.66
CardiacPCN [2]	-	17.19 ± 9.51	-	12.74 ± 6.59
Ours	**76.88 ± 4.54**	**14.97 ± 3.74**	**86.32 ± 5.11**	**9.98 ± 5.41**

artifacts, whereas the ones produced by SRHeart [19], MCSR-MS [4], and MCSR-ETE [5] are more motion-reduced. CardiacPCN [2] relies on post-processing and only generates low-quality meshes. In contrast, our method reconstructs surfaces that are more accurate and smooth.

3.3 Ablation Studies

The ablation studies are conducted in the clinical direct testing setting for our method, and the quantitative results are shown in Table 3. The shape accuracy of the super-resolved masks produced from MCSR-Net without further surface reconstruction using the latent diffusion is set as the baseline, which is lower than our proposed method. The next two configurations investigate different conditions, replacing C-SDGs with the super-resolved masks and the raw misaligned low-resolution masks, respectively. In these cases, the conditional encoders demand training from scratch with the diffusion models. Using low-resolution masks as conditions leads to significantly worse performances since it is more difficult to bridge them to the corresponding smooth surfaces, and utilizing super-resolved masks reaches a slightly decreased performance compared to employing C-SDGs. Besides, these configurations require the training of an additional conditional encoder, increasing computational cost and time.

We also explore the utilization of VAE [7] as the compression model, which leads to significantly inferior results. Given C-SDGs as conditions, the VQ-VAE encodes them into surface latent and replaces them with their vector-quantized latent from the codebook, which is trained using SDGs of high-fidelity surfaces. Therefore, even if C-SDGs are coarse and inaccurate, the vector-quantized latent suffers less from it. In contrast, VAE has weaker control over the latent space,

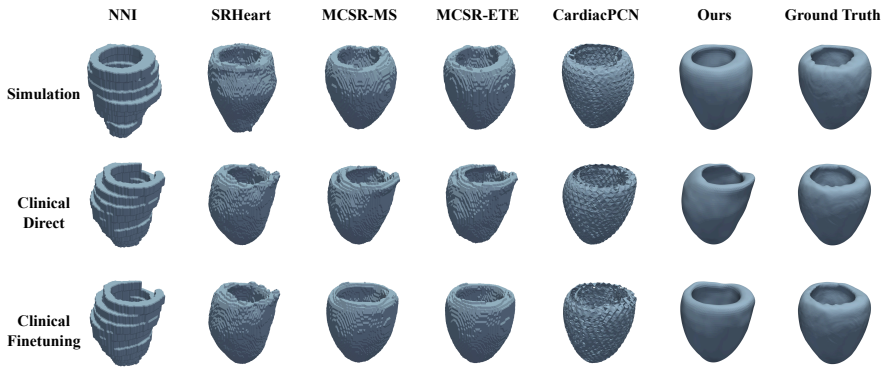

Fig. 2. The visualized reconstructed surfaces by the competing methods and ours.

yielding problematic latent representations when encountering poor-quality C-SDGs. In addition, we leverage a conditional GAN [10] to replace the latent diffusion to verify the functionality of the diffusion model. The result shows worse performances than our proposed method, demonstrating the effectiveness of the latent diffusion.

Table 3. The quantitative results of the ablation studies.

Configurations	Dice (%)	CD (mm)
MCSR-Net	73.04 ± 5.22	16.77 ± 3.92
Mask Conditioning	75.77 ± 4.68	15.54 ± 3.90
Low-Resolution Mask Conditioning	55.77 ± 11.37	44.08 ± 29.63
VAE [7]	55.63 ± 8.44	99.86 ± 313.54
Conditional GAN [10]	73.48 ± 5.72	17.03 ± 4.68
Proposed	**76.88 ± 4.54**	**14.97 ± 3.74**

4 Conclusion

In this paper, we propose a method for CMR motion correction and super-resolution to reconstruct high-fidelity cardiac surfaces based on latent diffusion models. Comprehensive experiments under both simulation and clinical settings have demonstrated that our method achieves the best surface reconstruction results both quantitatively and qualitatively. One potential aspect to improve is the design of MCSR-Net as the causes for motion artifacts are broader than slice-wise translations, which can serve as a direction for future work.

Acknowledgments. This work was supported by the ShanghaiTech AI4S Initiative (No. SHTAI4S202404) and the HPC Platform of ShanghaiTech University.

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

References

1. Bai, W., Shi, W., de Marvao, A., Dawes, T.J., O'Regan, D.P., Cook, S.A., Rueckert, D.: A bi-ventricular cardiac atlas built from 1000+ high resolution mr images of healthy subjects and an analysis of shape and motion. Medical image analysis **26**(1), 133–145 (2015)
2. Beetz, M., Banerjee, A., Grau, V.: Biventricular surface reconstruction from cine mri contours using point completion networks. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 105–109. IEEE (2021)
3. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
4. Chen, Z., Ren, H., Li, Q., Li, X.: Motion correction and super-resolution for multi-slice cardiac magnetic resonance imaging via a multi-stage deep learning approach. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–4. IEEE (2024)
5. Chen, Z., Ren, H., Li, Q., Li, X.: Motion correction and super-resolution for multi-slice cardiac magnetic resonance imaging via an end-to-end deep learning approach. Computerized Medical Imaging and Graphics **115**, 102389 (2024)
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
7. Kingma, D.P.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. 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)
10. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
11. 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)
12. Pinaya, W.H., Graham, M.S., Kerfoot, E., Tudosiu, P.D., Dafflon, J., Fernandez, V., Sanchez, P., Wolleb, J., Da Costa, P.F., Patel, A., et al.: Generative ai for medical imaging: extending the monai framework. arXiv preprint arXiv:2307.15208 (2023)
13. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3d deep learning with pytorch3d. arXiv preprint arXiv:2007.08501 (2020)
14. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
15. Savioli, N., de Marvao, A., O'Regan, D.: Cardiac super-resolution label maps. Mendeley Data (2021), v1, doi: 10.17632/pw87p286yx.1

16. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
17. Tarroni, G., Bai, W., Oktay, O., Schuh, A., Suzuki, H., Glocker, B., Matthews, P.M., Rueckert, D.: Large-scale quality control of cardiac imaging in population studies: application to uk biobank. Scientific reports **10**(1), 2408 (2020)
18. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
19. Wang, S., Qin, C., Savioli, N., Chen, C., O'Regan, D.P., Cook, S., Guo, Y., Rueckert, D., Bai, W.: Joint motion correction and super resolution for cardiac segmentation via latent optimisation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 14–24. Springer (2021)
20. Yuan, W., Khot, T., Held, D., Mertz, C., Hebert, M.: Pcn: Point completion network. In: 2018 international conference on 3D vision (3DV). pp. 728–737. IEEE (2018)
21. Zhou, Q.Y., Park, J., Koltun, V.: Open3d: A modern library for 3d data processing. arXiv preprint arXiv:1801.09847 (2018)