Proceedings Paper:
Zhang, Y. orcid.org/0009-0008-7725-6369, Chen, C. orcid.org/0000-0002-3525-9755, Shit, S. orcid.org/0000-0003-4435-7207 et al. (3 more authors) (2024) Whole heart 3D+T representation learning through sparse 2D cardiac MR images. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K. and Schnabel, J.A., (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. MICCAI 2024, 06-10 Oct 2024, Marrakesh, Morocco. Lecture Notes in Computer Science, 15001 . Springer Nature Switzerland , pp. 359-369. ISBN 9783031723773

https://doi.org/10.1007/978-3-031-72378-0_34

Whole Heart 3D+T Representation Learning Through Sparse 2D Cardiac MR Images

Yundi Zhang[1,2], Chen Chen[3,5,6], Suprosanna Shit[4], Sophie Starck[1,2], Daniel Rueckert[1,2,5], and Jiazhen Pan[1,2]

[1] School of Computation, Information and Technology, Technical University of Munich, Germany
[2] School of Medicine, Klinikum Rechts der Isar, Technical University of Munich, Germany
[3] Department of Computer Science, University of Sheffield, UK
[4] Department of Quantitative Biomedicine, University of Zurich, Switzerland,
[5] Department of Computing, Imperial College London, UK
[6] Department of Engineering Science, University of Oxford, UK
{yundi.zhang,jiazhen.pan}@tum.de

Abstract. Cardiac Magnetic Resonance (CMR) imaging serves as the gold-standard for evaluating cardiac morphology and function. Typically, a multi-view CMR stack, covering short-axis (SA) and 2/3/4-chamber long-axis (LA) views, is acquired for a thorough cardiac assessment. However, efficiently streamlining the complex, high-dimensional 3D+T CMR data and distilling compact, coherent representation remains a challenge. In this work, we introduce a whole-heart self-supervised learning framework that utilizes masked imaging modeling to automatically uncover the correlations between spatial and temporal patches throughout the cardiac stacks. This process facilitates the generation of meaningful and well-clustered heart representations without relying on the traditionally required, and often costly, labeled data. The learned heart representation can be directly used for various downstream tasks. Furthermore, our method demonstrates remarkable robustness, ensuring consistent representations even when certain CMR planes are missing/flawed. We train our model on 14,000 unlabeled CMR data from UK BioBank and evaluate it on 1,000 annotated data. The proposed method demonstrates superior performance to baselines in tasks that demand comprehensive 3D+T cardiac information, e.g. cardiac phenotype (ejection fraction and ventricle volume) prediction and multi-plane/multi-frame CMR segmentation, highlighting its effectiveness in extracting comprehensive cardiac features that are both anatomically and pathologically relevant.

1 Introduction

Cardiac Magnetic Resonance (CMR) imaging plays an essential role in the evidence-based diagnosis of cardiovascular diseases, establishing itself as the gold-standard [15] for cardiac morphology and function assessment by offering detailed 3D+T heart visualization. However, high-resolution 3D+T CMR acquisition is not practical in clinics due to the requirement for long breath-holds

and the reduced CMR contrast for further cardiac evaluation. Therefore, multi-view 2D+T CMR imaging including a stack of short-axis (SA) planes and 2/3/4 long-axis (LA) planes is preferable in clinical practice. Nonetheless, how to efficiently process these complex, high-dimensional CMR sequences and seamlessly integrate them into a coherent and unified 3D+T representation for a thorough cardiac assessment lacks a simple solution.

In the past decade, a large multitude of CMR representation learning methods have been proposed for cardiac function analysis [31, 2, 30, 23, 24]. They extract relevant features from CMR images, forming a compressed latent representation space essential for customized tasks. However, four major challenges remain in the field of cardiac representation learning. First, in most studies, the representation learning of cardiac morphology and function relies on curated annotated data. Yet, a vast majority of the CMR images are unlabeled. Only unsupervised/self-supervised (SSL) methods can leverage these large-scale CMR images and construct a comprehensive, meaningful representation. Second, most cardiac representation learning works focus on a specific downstream task, such as super-resolution and different modality mapping. Cardiac representation learning that can be generalized to multiple downstream tasks (e.g. cardiac segmentation and critical phenotype prediction) is still not broadly studied in the community. Third, existing studies demand a rigorous amount of inputs and lack adaptability and robustness to handle incomplete inputs, particularly when certain CMR planes are either not acquired or defective. Lastly, to the best of our knowledge, none of the previous studies efficiently incorporates all available spatial and temporal information from CMR. They either operate exclusively on a limited set of cardiac planes or neglect the utilization of temporal information.

In this paper, we introduce a cardiac representation learning method that is **self-supervised, scalable to vast unlabeled datasets, adaptable to diverse downstream tasks, flexible in handling varying amounts of input CMR planes, and capable of integrating comprehensive spatiotemporal information from sparse CMR inputs**. Our key contributions are:

1. We propose a 3D+T representation learning method for the whole heart, which is learned using multi-view (SA and LA) planes together with temporal information. By exploring correlation across different SA and LA sequences in an SSL manner, our model attains a rich and meaningful cardiac representation that can be adapted to different tasks.
2. Our approach ensures a consistent cardiac representation of the same subject, even in the absence of a few planes. This attribute is particularly beneficial in practice where some SA/LA planes are not available or of poor quality. This enables the same effectiveness as full CMR scans but with reduced acquisition costs and times.
3. We train our model with 14,000 unlabeled CMR data from UK-BioBank [21] and evaluate it on 1,000 annotated data. The visualized meaningful representation, the accurate cardiac phenotype prediction, e.g. ventricle volume and ejection fraction (which can only be achieved by leveraging adequate 3D+T information), and the enhanced end-to-end all-planes CMR segmentation

to baselines demonstrate its capability of dealing with various downstream tasks that require a comprehensive understanding of the whole heart. Its manifested versatility paves the way toward a cardiac MR foundation model.

1.1 Related Work

Cardiac imaging analysis can be conducted in various ways. Cardiac segmentation is the most common way to evaluate cardiac function. The relevant cardiac phenotype and clinic-useful parameters can be extracted based on it [14, 1, 6, 3]. However, whole-heart segmentation utilizing multi-view CMR sequences has not been widely studied. [5] used multi-view 2D CMR data to learn the correlation between different views and ameliorate segmentation performance but no temporal redundancy was exploited. [22] leveraged the temporal redundancy using RNNs and performed the 2D+T segmentation. [26] reconstructed high-resolution 3D SA segmentation using a neural implicit function but lacked LA knowledge integration. Moreover, critical cardiac phenotype/values can also be extracted directly from CMR without segmentation [33, 16, 31], which side-step the non-trivial manual cardiac imaging annotation. Cardiac motion/mesh tracking [29, 19, 17, 18] is another important approach to analyse cardiac morphology. Notably, both phenotype estimation (e.g. ejection fraction) and 3D cardiac motion require ample spatial/temporal information. Utilizing extensive 3D+T information from different views is thus advantageous for these analyses.

Cardiac representation learning simplifies the high-dimensional and complex CMR data into more manageable forms, enabling models to focus on the most relevant features of the heart. It can be applied to solve/facilitate different pre-defined tasks, e.g. cardiac segmentation [4, 2, 27], image reconstruction [25, 20], super-resolution [30] or cardiac indices prediction [31]. Notably, these methods predominantly relying on supervised learning face scalability issues in large unannotated datasets. Furthermore, cardiac representation can also bridge information from a different modality, e.g. ECG [28], genomics [24], and radiology reports [23], enhancing the breadth of data analysis. However, these approaches do not fully utilize the available 3D+T information from multi-view CMR, potentially missing critical spatial or temporal insights. Moreover, it is also challenging to align and correlate the information efficiently from these different views.

2 Methods

We develop our method's backbone based on Masked Autoencoders (MAE) [11] due to its alignment with our criteria, including inherent self-supervised learning characteristics, adaptability for various downstream tasks, and robustness in handling missing inputs. To incorporate high-dimensional cardiac information, we rebuild MAE from a simple 2D model to a cardiac representation learning method operating on complex multi-view CMR sequences. The training process

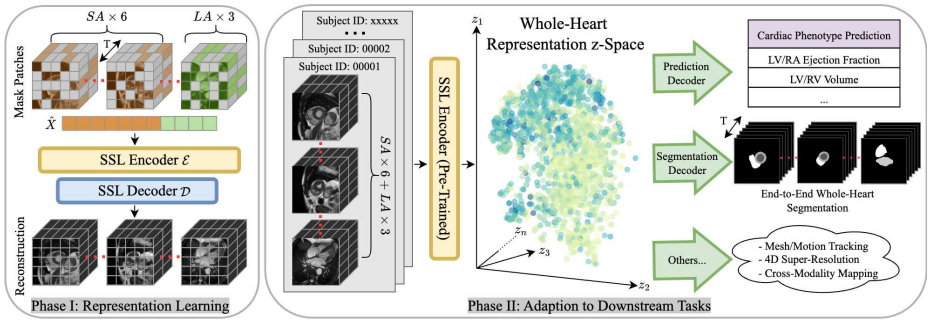

Fig. 1. Overview of the proposed method. Phase I: Representation learning is achieved through SSL reconstruction of a stack of multi-view masked 2D+T slices (6 SA and 3 LA). Phase II: Leveraging the whole-heart latent representation learned from the pre-trained SSL encoder, we utilize various decoders to carry out downstream tasks, e.g. cardiac phenotype prediction and whole-heart segmentation.

has two phases. In phase I, we leverage a huge amount of unlabeled data to learn representations. In phase II, the learned representations can be effortlessly extended to diverse downstream tasks, such as cardiac phenotype prediction and all-planes CMR segmentation. The model's architecture is illustrated in Fig. 1.

Phase I: Representation learning. We propose a simple but effective solution to correlate the "interlaced" multi-view CMR sequences and learn a whole-heart representation from them. We assume that each scan includes a set of sparse 2D+T CMR sequences with M SA planes $\{S_1, ..., S_M\}$ and N LA planes $\{L_1, ..., L_N\}$. Directly stacking all 2D+T sequences into a 4D tensor and conducting a 4D operation is computationally prohibitive. Nevertheless, LA and SA planes are acquired from different views and do not share spatial feature similarity. Thus, instead of forming a 4D tensor we first decompose each plane into P small 2D+T patches, denoted as s_p^m for the p-th patch of m-th SA plane and l_p^n for the p-th patch of n-th LA plane. Our method takes all these individual patches as an input vector, denoted as $X = [s_1^1, ..., s_P^M, l_1^1, ..., l_P^N]$, containing in total $(M + N) \times P$ patches. To capitalize on the temporal redundancy inherent in CMR data and further reduce computational demands, we use a larger patch size along the temporal than the spatial dimension. To enhance localization, each patch is added with a 4D positional embedding [11], indicating the patch's x–y–t spatiotemporal index and the corresponding plane of origin. Then a random mask is applied on the input to mask away q% of the patches. The shortened input with remaining patches, $\hat{X}$ (refer to Fig 1), are then fed into an encoder $\mathcal{E}$ to learn a dense 3D+T cardiac representation. A high maskout ratio q% is chosen here to force the encoder to find the underlying spatiotemporal correlation across different cardiac planes, rather than easily extrapolate missing pixel intensity. Afterwards, a lightweight decoder $\mathcal{D}$ is applied to predict the masked-

out patches and reconstruct X. We applied the mean squared error (MSE) for optimization with a loss function formulated as $\mathcal{L} = \left\| X - \mathcal{D}\left(\mathcal{E}\left(\hat{X}\right)\right)\right\|_2^2$.

Phase II: Adaption to different downstream tasks. The 3D+T whole heart representation can be used to solve different downstream tasks. We finetune our model on a small-scale dataset where corresponding labels are available. In phase II, we forward all available CMR planes to the model without maskout. The encoder remains unchanged and the reconstruction decoder is replaced with a different decoder according to the task we aim to solve. For cardiac phenotype (e.g. ejection fraction and volume) prediction, we use shallow linear layers since we assume the spatiotemporal representation learned in phase I is already comprehensive and can represent cardiac morphology to some extent. For all-planes CMR sequences segmentation, a U-Net-wise decoder with skip-connections [34] is adopted. It is noteworthy that, in contrast to conventional cardiac segmentation (refer to section 1.1) which can only conduct specific 2D plane segmentation at one inference time, our model can segment **all CMR planes/frames** in an end-to-end fashion at a **single** inference time. Furthermore, since the model can exploit the correlation between different CMR planes, the final segmentation is further enhanced and is expected to perform better than single plane segmentation [5] (more details refer to Table 2 and Fig. 3).

Robustness when some planes are absent. In clinic scenarios, variations in acquisition protocols or artifacts from patient motion can lead to missing/corrupted CMR planes. Our method, however, remains robust against these challenges due to its contrastive learning inherence. Contrastive learning is usually used in representation learning to pull positive (similar) data and push dissimilar data [9, 7]. Our method employs a form of implicit contrastive learning, aligning positive pairs through diverse mask patterns [32]. During different training epochs in phase I, our model is forced to generate the same representation/reconstruction for a subject, regardless of different random mask applications. This ensures that even in scenarios where certain CMR planes are missing, our model can still deliver the same representation as complete CMR plane inputs.

3 Dataset and Experiments

Dataset. Our model is trained on the CMR data from UK BioBank [21]. We use 14,000 subjects to conduct representation learning and finetune the model on different downstream tasks using 1,000 annotated subjects with segmentation masks and cardiac phenotype labels from [1]. The tests are performed on 100 subjects. Each subject in the dataset includes 6 SA and 3 LA 2D slices with 50 time frames. All slices are cropped to the cardiac center of size 128×128.
Implementation details. We implement 6 encoder layers and 2 decoder layers in SSL with an embedding dimension of 1024. The patch size is $8 \times 8 \times 25$ with 8 the spatial and 25 the temporal dimension. The mask ratio $q\%$ in SSL is set to 70%. For downstream tasks, we set the embedding dimension to 256 for the

prediction decoder and 576 for the segmentation decoder. We conduct all experiments on a NVIDIA A6000 GPU. The batch size is set to 1 for representation learning and segmentation, and 4 for cardiac phenotype prediction. The initial learning rate (LR) is set to 10^{-4} and we use a cosine scheduler with a weight decay of 0.05. The LR for downstream tasks' fine-tuning is set to 10^{-5}.

Downstream baseline methods. For prediction, we compare our method with ResNet50 [12] and ViT [8]. They both take concatenated 3D+T planes as input and T is treated as channel dimension. For multi-plane segmentation, we compare our method with nnUNet [13] and UNETR [10]. As nnUNet can only take 2D+T planes as inputs, we have to carry out two times training (once for SA and once for LA) and treat different planes separately. We adapt UNETR to UNETR+ so that it can take all 2D+T planes as input at once (therefore also 3D+T), aligning with our setting. Moreover, we compare to our ablated methods which are trained only using LA or SA CMR. This allows us to assess the extent to which performance can be enhanced when additional spatiotemporal information from other planes is incorporated.

4 Results and Discussion

Representation learning evaluation via reconstruction. We first evaluate reconstruction using only SA sequences, only LA sequences, and all available views in pre-training. The quantitative results and reconstructed CMR images are shown in Appendix Table 3 and Fig. 4. The superior results of all-view reconstruction to that of using only limited spatial planes imply the benefit of involving all-view spatiotemporal information for whole-heart representation learning. Incorporating different views can help the model build spatial knowledge of the heart and therefore enhance the representation/reconstruction.

Representation learning evaluation via t-SNE visualization. We further provide the t-SNE visualization of the learned representations in Fig. 2. We label the latent embeddings after pre-training with right ventricular ejection fraction (RVEF) and left ventricle mass (LVM). The generation of these two phenotypes relies on 3D spatial and temporal information of the heart. Notably, even without using any labels during pre-training, our model can already generate a well-clustered latent space reflective of spatiotemporal differences across the subjects. This visualization serves as strong evidence for the effectiveness of the learned representations.

Cardiac phenotype prediction. We assess the efficacy of learned whole-heart representations by predicting critical phenotype values. Table 1 shows the mean absolute error of our predictions compared to ResNet50 and ViT for LVM, RVEF, right atrial ejection fraction (RAEF), right ventricular end-diastolic volume (RVEDV), and left atrial stroke volume (LASV). For accurate prediction of these values, both adequate spatial and temporal cardiac information is essential. The superior performance to the baseline methods underscores its capability to capture high-level spatiotemporal features of the entire heart.

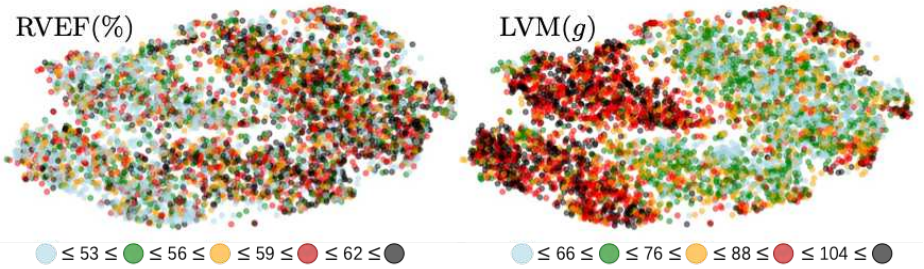

Fig. 2. The t-SNE visualization of the latent whole-heart representation obtained through pre-training. Latent embeddings are labeled with RVEF and LVM values, categorized into 5 groups according to the ground truth, and shown in different colors.

Table 1. Comparison of mean absolute errors among mean-guess (estimating every subject's phenotype value with the cohort mean value), ResNet50, ViT, and the proposed approach for predicting LVM, RVEF, RAEF, RVEDV, and LASV. ResNet50 and ViT use concatenated 3D+T planes as input.

Method	LVM (g)	RVEF (%)	RAEF (%)	RVEDV (mL)	LASV (mL)
Mean-guess	$17.103_{\pm 11.104}$	$12.14_{\pm 6.808}$	$6.586_{\pm 5.899}$	$31.915_{\pm 22.807}$	$8.918_{\pm 6.113}$
ResNet50	$7.150_{\pm 6.945}$	$3.245_{\pm 2.354}$	$6.294_{\pm 4.983}$	$12.577_{\pm 11.003}$	$5.777_{\pm 4.695}$
ViT	$8.656_{\pm 7.858}$	$4.292_{\pm 3.197}$	$6.590_{\pm 5.486}$	$14.102_{\pm 12.758}$	$8.944_{\pm 6.219}$
Ours	$\mathbf{4.332_{\pm 4.716}}$	$\mathbf{2.831_{\pm 2.550}}$	$\mathbf{5.396_{\pm 4.213}}$	$\mathbf{10.713_{\pm 8.980}}$	$\mathbf{3.529_{\pm 3.186}}$

Multi-view segmentation. The competence of our model to perform end-to-end segmentation across all planes is shown in Table 2 and Fig. 3. Our model not only exhibits comparable quantitative dice scores with nnUNet (that leverages exhaustive parameters tuning) but also shows superior performance over UNETR+ in all regions for both SA/LA planes. This benefit is gained from the learned whole-heart representation. Moreover, the superior performance against SA-only and LA-only segmentation also highlights the significance of integrating multi-view CMR information for more accurate segmentation outcomes.

Representation learning robustness. To simulate scenarios with incomplete or defective CMR data, we randomly remove 1 or 2 CMR planes from each subject and generate a new latent representation. Calculating the cosine similarity between this representation and the one with all planes available, we observe an average value of 1.0 in both cases, indicating a total alignment. Further, we conduct downstream LVM and RVEF predictions with 2 slices masked out and compare them with the regular setting (no maskout). The predicted value difference is ±0.3g for LVM and ±0.4% for RVEF, emphasizing its robustness and significant potential in practical scenarios with missing planes.

Table 2. Segmentation dice scores of nnUNet, UNETR+, and proposed methods trained with SA-only, LA-only and all-planes CMR. nnUNet employs a single 2D+T plane as input (therefore 2 times training for SA/LA) while UNETR+ uses all sparse CMR sequences as input, same as ours. The top two results are marked in bold.

Phenotype	nnUNet	UNETR+	Ours (SA)	Ours (LA)	Ours (All)
LVBP	**0.98**$_{\pm 0.01}$	0.89$_{\pm 0.02}$	0.95$_{\pm 0.02}$	N.A.	**0.96**$_{\pm 0.01}$
LVMYO	**0.96**$_{\pm 0.02}$	0.85$_{\pm 0.04}$	0.81$_{\pm 0.04}$	N.A.	**0.89**$_{\pm 0.02}$
RVBP	**0.97**$_{\pm 0.02}$	0.88$_{\pm 0.03}$	0.91$_{\pm 0.04}$	N.A.	**0.91**$_{\pm 0.02}$
LABP	**0.96**$_{\pm 0.02}$	0.92$_{\pm 0.03}$	N.A.	0.93$_{\pm 0.03}$	**0.94**$_{\pm 0.02}$
RABP	**0.98**$_{\pm 0.03}$	0.93$_{\pm 0.05}$	N.A.	0.93$_{\pm 0.06}$	**0.94**$_{\pm 0.04}$

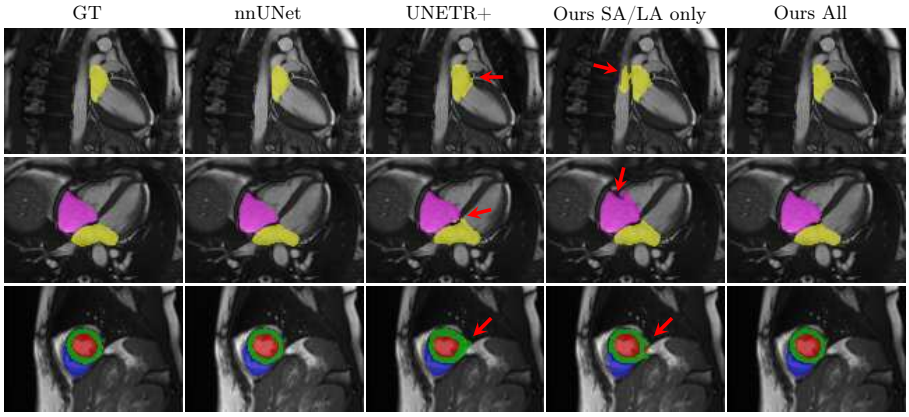

Fig. 3. Qualitative segmentation results among nnUNet, UNETR+, and the proposed methods. UNETR+ and the proposed approach in the last column (Ours All) use all sparse CMR sequences as network input, while nnUNet and the second last column (Ours SA/LA) are trained solely with either SA or LA views.

Limitation and outlook. While the presented evaluation is limited to prediction and segmentation tasks, future work will further explore tasks requiring whole 3D+T information, such as cardiac motion tracking and other modality mapping (e.g. genetics) to widen the scope of cardiac analysis. While our model demonstrates meaningful representations across diverse subjects, it has not been evaluated on longitudinal data. This presents an avenue for tracking potential cardiac disease progression by observing patient trajectories in the latent space. Additionally, the capability of our method to offer comprehensive cardiac representations with fewer CMR images opens opportunities to reduce CMR scan times and alleviate patient discomfort.

5 Conclusion

In this study, we introduced a self-supervised method for 3D+T cardiac representation learning, utilizing multi-view sparse 2D CMR images. Trained on 14,000 CMR datasets, our approach is adaptable to various downstream tasks and maintains robust performance even with missing CMR slices. Leveraging the heart's spatiotemporal information, our model enables accurate cardiac phenotype prediction and efficient, precise whole-heart segmentation.

6 Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 87802. This work is funded by the European Research Council (ERC) project Deep4MI (884622).

References

1. Bai, W., Suzuki, H., Huang, J., Francis, C., et al.: A population-based phenome-wide association study of cardiac and aortic structure and function. Nature medicine **26**(10), 1654–1662 (2020)
2. Biffi, C., Oktay, O., Tarroni, G., Bai, W., et al.: Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling. In: MICCAI. pp. 464–471 (2018)
3. Campello, V.M., Gkontra, P., Izquierdo, C., Martin-Isla, C., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. IEEE Transactions on Medical Imaging **40**(12), 3543–3554 (2021)
4. Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., et al.: Factorised spatial representation learning: Application in semi-supervised myocardial segmentation. In: MICCAI. pp. 490–498 (2018)
5. Chen, C., Biffi, C., Tarroni, G., Petersen, S., et al.: Learning shape priors for robust cardiac mr segmentation from multi-view images. In: MICCAI. pp. 523–531. Springer (2019)
6. Chen, C., Qin, C., Qiu, H., Tarroni, G., et al.: Deep learning for cardiac image segmentation: a review. Frontiers in Cardiovascular Medicine **7**, 25 (2020)
7. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR. pp. 15750–15758 (2021)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Grill, J.B., Strub, F., Altché, F., Tallec, C., et al.: Bootstrap your own latent-a new approach to self-supervised learning. NeurIPS **33**, 21271–21284 (2020)
10. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., et al.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)

13. 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)
14. Khened, M., Kollerathu, V.A., Krishnamurthi, G.: Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. Medical image analysis **51**, 21–45 (2019)
15. von Knobelsdorff-Brenkenhoff, F., Pilz, G., Schulz-Menger, J.: Representation of cardiovascular magnetic resonance in the aha/acc guidelines. Journal of Cardiovascular Magnetic Resonance **19**(1), 1–21 (2017)
16. Luo, G., Sun, G., Wang, K., Dong, S., Zhang, H.: A novel left ventricular volumes prediction method based on deep learning network in cardiac mri. In: Computing in Cardiology Conference. pp. 89–92 (2016)
17. Meng, Q., Qin, C., Bai, W., Liu, T., et al.: Mulvimotion: Shape-aware 3d myocardial motion tracking from multi-view cardiac mri. IEEE TMI **41**(8), 1961–1974 (2022)
18. Pan, J., Huang, W., Rueckert, D., Küstner, T., Hammernik, K.: Reconstruction-driven motion estimation for motion-compensated mr cine imaging. IEEE TMI (2024)
19. Pan, J., Rueckert, D., Küstner, T., Hammernik, K.: Efficient image registration network for non-rigid cardiac motion estimation. In: MICCAI. pp. 14–24 (2021)
20. Pan, J., Shit, S., Turgut, Ö., Huang, W., et al.: Global k-space interpolation for dynamic mri reconstruction using masked image modeling. In: MICCAI. pp. 228–238 (2023)
21. Petersen, S.E., Matthews, P.M., Francis, J.M., Robson, M.D., et al.: UK Biobank's cardiovascular magnetic resonance protocol. JCMR pp. 1–7 (2015)
22. Qin, C., Bai, W., Schlemper, J., Petersen, S.E., et al.: Joint learning of motion estimation and segmentation for cardiac mr image sequences. In: MICCAI. pp. 472–480. Springer (2018)
23. Qiu, J., Huang, P., Nakashima, M., Lee, J., et al.: Multimodal representation learning of cardiovascular magnetic resonance imaging. arXiv preprint arXiv:2304.07675 (2023)
24. Radhakrishnan, A., Friedman, S.F., Khurshid, S., Ng, K., et al.: Cross-modal autoencoder framework learns holistic representations of cardiovascular state. Nature Communications **14**(1), 2436 (2023)
25. Schlemper, J., Oktay, O., Bai, W., Castro, D.C., et al.: Cardiac mr segmentation from undersampled k-space using deep latent representation learning. In: MICCAI. pp. 259–267 (2018)
26. Stolt-Ansó, N., McGinnis, J., Pan, J., Hammernik, K., Rueckert, D.: Nisf: Neural implicit segmentation functions. In: MICCAI. pp. 734–744. Springer (2023)
27. Sun, X., Liu, Z., Zheng, S., Lin, C., et al.: Attention-enhanced disentangled representation learning for unsupervised domain adaptation in cardiac segmentation. In: MICCAI. pp. 745–754 (2022)
28. Turgut, Ö., Müller, P., Hager, P., Shit, S.e.a.: Unlocking the diagnostic potential of ecg through knowledge transfer from cardiac mri. arXiv preprint arXiv:2308.05764 (2023)
29. Wang, H., Amini, A.A.: Cardiac motion and deformation recovery from mri: a review. IEEE TMI **31**(2), 487–503 (2011)
30. Wang, S., Qin, C., Savioli, N., Chen, C., et al.: Joint motion correction and super resolution for cardiac segmentation via latent optimisation. In: MICCAI. pp. 14–24 (2021)

31. Xue, W., Islam, A., Bhaduri, M., Li, S.: Direct multitype cardiac indices estimation via joint representation and regression learning. IEEE TMI **36**(10), 2057–2067 (2017)
32. Zhang, Q., Wang, Y., Wang, Y.: How mask matters: Towards theoretical understandings of masked autoencoders. NeurIPS **35**, 27127–27139 (2022)
33. Zhen, X., Wang, Z., Islam, A., Bhaduri, M., et al.: Multi-scale deep networks and regression forests for direct bi-ventricular volume estimation. Medical image analysis **30**, 120–129 (2016)
34. Zhou, L., Liu, H., Bae, J., He, J., et al.: Self pre-training with masked autoencoders for medical image classification and segmentation. In: IEEE ISBI. pp. 1–6 (2023)

A Supplementary

Table 3. Reconstruction PSNR (dB) for SA and LA views. From left to right it shows the PSNR of our model trained with only SA 2D+T planes, only LA 2D+T planes, and all available multi-view planes separately.

Eval. \ Input	SA	LA	ALL
SA	$26.20_{\pm 0.84}$	N.A.	$\mathbf{31.08_{\pm 0.86}}$
LA	N.A.	$25.11_{\pm 1.48}$	$\mathbf{28.60_{\pm 1.50}}$

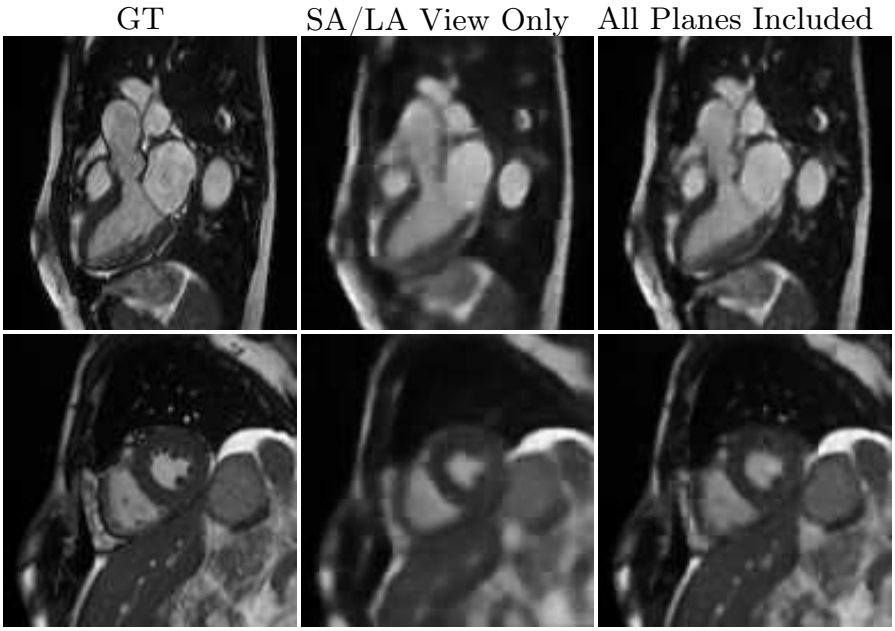

Fig. 4. Reconstruction samples in the pre-training phase. The second column shows reconstructions with input planes from a single view (SA/LA only), while the last column shows the reconstructions with planes from both views included.