Geometry-Aware Self-Supervised Learning for Whole-Heart 3D+t Representation

Liwei Hu[1], Zi Wang[1], Yinzhe Wu[1], Zhenxuan Zhang[1], Anbang Wang[1], Ruicheng Yuan[1], Fanwen Wang[1], Choon Hwai Yap[2✉], and Guang Yang[1,3,4,5✉]

[1] Bioengineering Department and Imperial-X, Imperial College London, London, UK
[2] Department of Bioengineering, Imperial College London, London, UK
[3] National Heart and Lung Institute, Imperial College London, London, UK
[4] Cardiovascular Research Centre, Royal Brompton Hospital, London, UK
[5] School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK
{c.yap,g.yang}@imperial.ac.uk

Abstract. Cardiac magnetic resonance (CMR) is a clinical reference standard for assessing cardiac morphology and function. Whole-heart assessment relies on sparse multi-plane cine acquisitions from complementary short-axis (SAX) and 2/3/4-chamber long-axis (LAX) views across the cardiac cycle. Although large-scale cohorts such as UK Biobank provide tens of thousands of such studies, most scans remain unlabeled, and learning compact whole-heart representations from the resulting high-dimensional 3D+t data is still challenging under practical memory constraints. To address this, we introduce a geometry-aware self-supervised learning framework for whole-heart representation learning from sparse multi-plane cine CMR. First, we inject patient-space coordinates into spatio-temporal tubelet tokens, which preserves explicit spatial identity and enables shared per-plane encoding. Next, we explicitly model cross-plane structure by aggregating multi-plane features into a latent 3D volume via voxel-query cross-attention. To ensure geometric consistency while keeping supervision efficient, we reconstruct masked tubelets by coordinate-constrained trilinear reprojection from the latent volume. Finally, we adopt hierarchical masking that combines intra-plane tubelet masking with inter-plane plane dropping, turning reconstruction into cross-plane completion and strengthening global 3D reasoning. Experiments show that coherent 3D anatomy can be decoded from the learned latent volume using a lightweight linear probe, and that masked-plane reconstruction recovers missing views from remaining planes. Moreover, our method improves downstream multi-plane segmentation over state-of-the-art methods, while maintaining favorable efficiency under the same input setting. Code is available at this URL.

Keywords: Geometry-aware modelling · Self-supervised Learning · Cardiac MRI · Multi-plane.

1 Introduction

Cardiac magnetic resonance (CMR) imaging is widely used for assessing cardiac structure and function [1,12,20,21]. Large-scale cohorts such as UK Biobank (UKB) [13] have collected tens of thousands of standardized multi-plane cine CMR studies from complementary short-axis (SAX) and long-axis (LAX) views across the cardiac cycle. These data offer rich 3D+t spatio-temporal information, but expert annotations remain scarce. Self-supervised learning (SSL) provides a principled way to leverage such unlabeled cohorts by learning strong, transferable representations directly from the raw imaging signals [3,8,14,17].

Despite its promise, SSL for whole-heart cine CMR is non-trivial. Cine CMR is inherently high-dimensional and multi-plane, forming a large 3D+t signal across SAX and LAX views. Directly applying dense 3D+t SSL is often prohibitively expensive in memory and computation, so many existing approaches adopt simplified training protocols, such as using partial planes, reduced temporal coverage, or heavily downsampled inputs [4,5,10]. As a consequence, substantial spatio-temporal information is discarded, which can limit the learned representation's ability to capture coherent whole-heart geometry and motion. Recent works further adapt video masked autoencoders by patchifying each plane and concatenating tokens from multiple planes into a single sequence, using view-aware positional embeddings to encourage cross-view interaction [14,24]. However, such implicit alignment remains challenging because orthogonal SAX and LAX planes intersect only sparsely and may exhibit large out-of-plane gaps. Masked reconstruction can admit shortcut solutions that exploit local temporal redundancy or intra-plane inpainting rather than learning explicit 3D structure.

To address these challenges, we propose a geometry-aware SSL framework that avoids treating cine CMR as a dense 3D+t tensor and instead injects patient-space physical coordinates to make cross-plane relations explicit in a shared reference frame, thereby simplifying cross-view modelling and improving pretraining efficiency. Concretely, we augment each spatio-temporal tubelet token with a DICOM-derived patient-space coordinate and encode each plane independently using a shared encoder, preserving explicit spatial identity across views. We then model cross-plane structure explicitly by aggregating the encoded multi-plane features into a latent 3D volume via voxel-query cross-attention, and enforce geometric consistency by coordinate-constrained trilinear reprojection for masked reconstruction. To further suppress intra-plane shortcut solutions, we adopt hierarchical masking that combines intra-plane tubelet masking with inter-plane masking that drops entire planes, turning reconstruction into cross-plane completion and strengthening global 3D reasoning. Our contributions are three-fold:

1. We propose a geometry-aware shared encoding strategy that supports sparse and unordered multi-plane cine CMR via patient-space coordinate injection.
2. We develop an explicit plane-to-volume modelling and coordinate-based reprojection mechanism for geometry-consistent self-supervised pretraining.

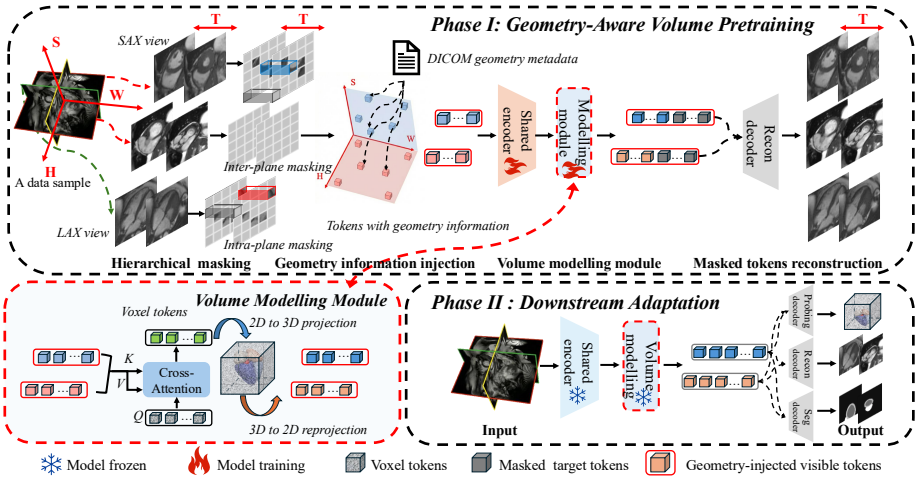


Fig. 1: Overview of the proposed two-phase framework. **Phase I** performs geometry-aware volume pretraining via self-supervised masked tubelet reconstruction on sparse multi-plane cine CMR. Token colors indicate different input views, while colored borders denote geometry-injected visible tokens. Masked target tokens are used only as reconstruction targets. **Phase II** freezes the pretrained backbone and trains lightweight heads for masked-plane reconstruction, latent volume probing, and multi-plane segmentation.

3. We conduct comprehensive evaluations on UKB, including latent volume probing, masked-plane reconstruction, multi-plane segmentation, and efficiency analysis against strong baselines.

2 Methodology

As shown in Fig. 1, Phase I learns geometry-aware representations via self-supervised masked reconstruction through a latent 3D volume, and Phase II evaluates the learned representation on downstream evaluations, including latent volume probing and multi-plane segmentation.

2.1 Problem Formulation

We adopt masked spatio-temporal tubelet reconstruction as our pretraining objective [6,16]. In cine CMR, each scan is an unordered set of $S = M + K$ sparse 2D+t plane sequences $\mathcal{X} = \{x_i\}_{i=1}^{S}$, where $x_i \in \mathbb{R}^{T\times H\times W}$. We partition each x_i into spatio-temporal tubelets $p_{i,j}$ of size $\tau \times P \times P$. For clarity, we call $p_{i,j}$ tubelets, and refer to their embedded representations as tubelet tokens.

Self-supervised Objective. Let Ω denote the index set of masked tubelets, and let $\mathcal{X}_{\text{vis}} = \{p_{i,j} \mid (i,j) \notin \Omega\}$ be the remaining visible tubelets. Compared

with natural videos, cine CMR consists of sparse multi-plane sequences whose planes intersect only partially, making overlap-driven fusion unreliable [15,22]. We therefore abstract cross-plane interaction as a fusion operator $\mathcal{F}$ and adopt the asymmetric masked autoencoding formulation [6,16]: the encoder Φ_{enc} is applied only to visible tubelet tokens, and the lightweight decoder Φ_{dec} reconstructs masked targets from the fused representation. For each masked tubelet $(i,j) \in \Omega$,

$$\hat{p}_{i,j} = \Phi_{\text{dec}}\Big(\mathcal{F}(\mathcal{X}_{\text{vis}});\ i,j\Big), \quad (i,j) \in \Omega. \tag{1}$$

We optimize the reconstruction loss over masked tubelets:

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \|\hat{p}_{i,j} - p_{i,j}\|_2^2 . \tag{2}$$

2.2 Phase I: Geometry-Aware Volume Pretraining

A straightforward implementation of the fusion operator $\mathcal{F}$ concatenates tubelet tokens from all planes and applies a joint transformer [14,24], where self-attention scales quadratically with the total token count, i.e., $\mathcal{O}(N^2)$. We instead inject patient-space geometry to enable shared per-plane encoding and realize $\mathcal{F}$ with plane-to-volume fusion $\mathcal{R}$ and volume-to-plane reprojection $\mathcal{G}$.

Geometry-aware Shared Encoding. We augment each tubelet with a patient-space coordinate derived from DICOM geometry metadata. Specifically, for plane sequence x_i, let $\mathbf{o}_i \in \mathbb{R}^3$ be `ImagePositionPatient`, let $\mathbf{r}_i, \mathbf{s}_i \in \mathbb{R}^3$ be the row/column direction cosines from `ImageOrientationPatient`, and let (Δ_u, Δ_v) be `PixelSpacing`. For a tubelet $p_{i,j}$ centered at pixel location $(u_{i,j}, v_{i,j})$, its 3D patient-space coordinate is:

$$\mathbf{c}_{i,j} = \mathbf{o}_i + (u_{i,j}\Delta_u)\mathbf{r}_i + (v_{i,j}\Delta_v)\mathbf{s}_i. \tag{3}$$

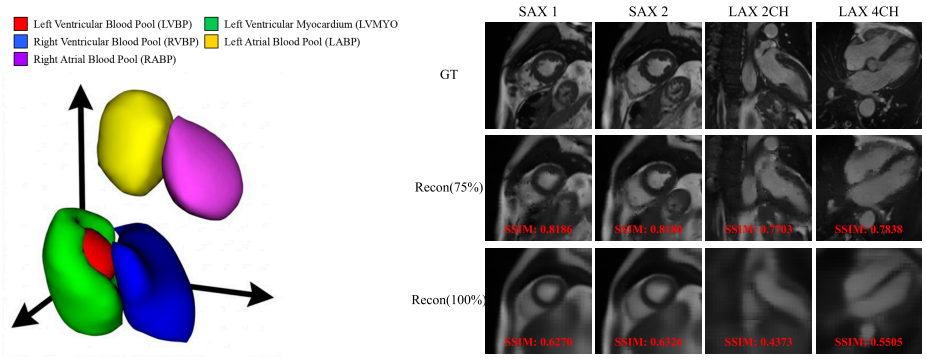
We encode $\mathbf{c}_{i,j}$ with an MLP and add it to the tubelet content embedding,

$$h_{i,j} = e_{\text{tube}}(\mathbf{p}_{i,j}) + e_{\text{geo}}(\mathbf{c}_{i,j}), \quad (i,j) \notin \Omega. \tag{4}$$

where $e_{\text{tube}}(\cdot)$ denotes the tubelet content embedding layer, and $e_{\text{geo}}(\cdot)$ is implemented as an MLP projecting $\mathbf{c}_{i,j}$ to the token dimension. Following asymmetric masked autoencoding, we apply the shared encoder only to visible geometry-tagged tubelet tokens, producing $\tilde{h}_{i,j} = \Phi_{\text{enc}}(h_{i,j})$ for $(i,j) \notin \Omega$, and pass $\tilde{h}_{i,j}$ to the volume modelling module.

Complexity discussion. Let each plane yield n tubelet tokens and let S be the number of planes ($N = Sn$). Early fusion applies self-attention over all tokens, giving $\mathcal{O}(N^2) = \mathcal{O}(S^2n^2)$ complexity. With shared per-plane encoding, the same encoder processes planes independently, resulting in $\mathcal{O}(S\,n^2)$ for encoding, while cross-plane interaction is deferred to the plane-to-volume module.

Plane-volume Fusion and Reprojection. We implement the operator $\mathcal{F}$ as a two-step pathway: plane-to-volume fusion $\mathcal{R}$ followed by volume-to-plane reprojection $\mathcal{G}$.



(a) (b)

Fig. 2: Representation analyses of the learned latent volume. (a) Latent volume probing using a frozen backbone and a linear probe to decode coarse 3D anatomy on a dense coordinate grid. (b) Masked-plane reconstruction under 75% tubelet masking and full plane dropping (100% masking) with SSIM.

Plane-to-volume fusion. To explicitly model cross-plane relations, we fuse the visible multi-plane features into a latent volume V. Concretely, learnable voxel queries Q aggregate the encoded multi-plane features $\tilde{H}$ via cross-attention, producing voxel tokens [11,23]. We then reshape the resulting tokens into a low-resolution voxel lattice to form V:

$$V = \mathcal{R}(\tilde{H}) = \text{reshape}\Big(\text{Attn}(W_q Q,\ W_k \tilde{H},\ W_v \tilde{H})\Big)\,. \tag{5}$$

Volume-to-plane reprojection. Multi-plane cine sequences provide robust supervision, since cardiac shape and motion cues are consistently observed over time within each plane. We therefore compute the reconstruction loss in tubelet space by reprojection from V, which is also more efficient than volumetric supervision. Instead of a learned reprojection module, we define the volume-to-plane reprojection operator $\mathcal{G}$ as coordinate-constrained trilinear sampling, enforcing a strong geometric prior:

$$\hat{p}_{i,j} = \Phi_{\text{dec}}(\mathcal{G}(V, \mathbf{c}_{i,j}))\,, \quad (i,j) \in \Omega. \tag{6}$$

Hierarchical Masking for Shape Completion. With only intra-plane tubelet masking, reconstruction may degenerate into plane-wise inpainting after reprojection, leaving the latent volume under-constrained. We therefore introduce an inter-plane masking scheme in which, at each iteration, one entire plane sequence is randomly dropped, so reconstruction must rely on cross-plane completion through the shared latent volume. Recalling that Ω denotes the index set of masked tubelets, we decompose it as:

$$\Omega = \Omega_{\text{intra}} \cup \Omega_{\text{inter}}, \qquad \Omega_{\text{inter}} = \{(i,j) \mid i \in \mathcal{S}_{\text{mask}}\}, \tag{7}$$

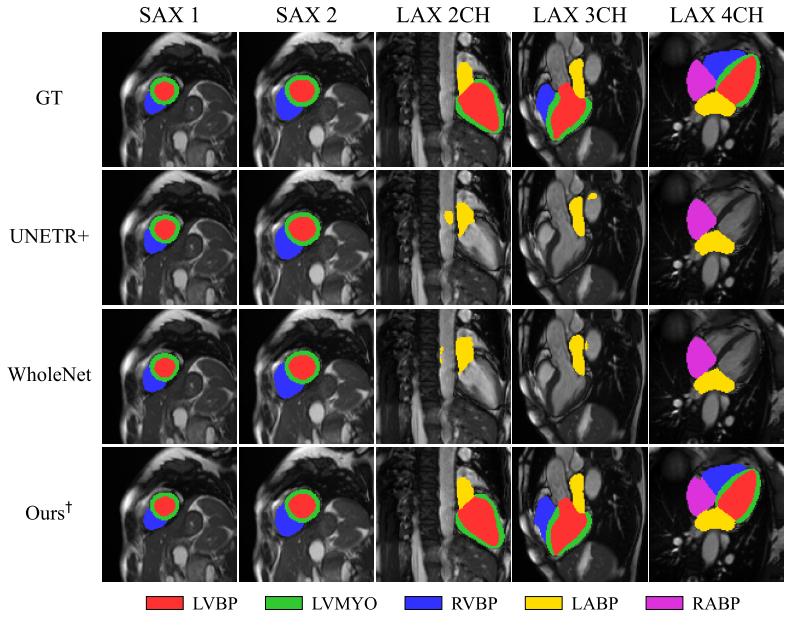


Fig. 3: Qualitative multi-plane segmentation results. Columns correspond to cine planes (SAX mid/basal slices and LAX 2CH/3CH/4CH views). Rows show ground truth (GT) and predictions from UNETR+ [7], WholeNet [24], and Ours† (compared baselines only output view-specific labels on LAX views).

where Ω_{intra} is the tubelet mask set on visible planes and $\mathcal{S}_{\text{mask}} \subseteq \{1, \ldots, S\}$ denotes the set of dropped planes. Accordingly, the visible tubelets are $\mathcal{X}_{\text{vis}} = \{p_{i,j} \mid i \notin \mathcal{S}_{\text{mask}}, (i,j) \notin \Omega_{\text{intra}}\}$, which compels the model to reconstruct the dropped plane via cross-plane fusion in the latent volume under the same objective in Eq. (2).

2.3 Phase II: Downstream Adaptation

We freeze the pretrained backbone and attach lightweight heads for downstream evaluation.

Representation analyses. We evaluate the learned 3D structure via (i) masked-plane reconstruction, which drops an entire plane and reconstructs it using the pretrained reconstruction path, and (ii) latent volume probing, which trains a linear probe on the low-resolution latent volume.

Multi-plane segmentation. We train a segmentation head on top of the frozen backbone and report multi-plane segmentation results. The backbone supports variable numbers of input planes at inference, enabling evaluation under sparse inputs.

3 Experiment Setting

Dataset and Preprocessing. We use cine cardiac MR from UK Biobank [13]. We pretrain on 4,000 subjects, evaluate on 1,000 annotated subjects with segmentation masks from [18], and test on 100 held-out subjects. Each subject contains 9 cine planes (6 SAX + 3 LAX) with 50 frames; all planes are center-cropped to 112×112 pixels and temporally sampled to $T = 32$ frames. Unless otherwise specified, we use $|\mathcal{D}_{\text{pre}}| = 4\text{k}$ for pretraining and $|\mathcal{D}_{\text{seg}}| = 1\text{k}$ for segmentation.

Implementation Details. We use an asymmetric architecture with a ViT encoder (L=9, D=192, H=3) and a 3-layer decoder (D=96, H=3) [19]. Inputs are tokenized into tubelets $[t, h, w] = [8, 4, 4]$, and the volume modelling module maintains a 16^3 latent grid. All experiments run on a single NVIDIA RTX PRO 6000 GPU using AdamW (lr $= 10^{-4}$, weight decay $= 0.05$) with cosine scheduling for 300 epochs and 20 warm-up epochs. For downstream evaluation, we freeze the pretrained backbone and assess it with two representation analyses (masked-plane reconstruction using the pretrained reconstruction decoder, and latent volume probing via a linear probe), and with a downstream multi-plane segmentation head (lr $= 10^{-4}$).

Baselines and Ablations. We compare against nnUNet [9], UNETR [7], and WholeNet [24]. nnUNet is trained separately for SAX and LAX due to its single-plane input, while UNETR and WholeNet follow the same multi-plane input protocol as ours.

4 Results and Discussion

Representation Analyses. We evaluate the capacity of the pretrained backbone to capture cohesive 3D structures using two representation analyses. (i) Latent volume probing (Fig. 2a) freezes the backbone and volume modelling module, and trains only a lightweight linear probe on the latent volume. By querying this representation with a dense isotropic coordinate grid (e.g., at 0.2 mm resolution), the simple probe recovers coarse but structured 3D cardiac geometry from the latent volume. This confirms that spatially organized anatomical information is embedded within the learned representation. (ii) Masked-plane reconstruction (Fig. 2b) evaluates cross-plane completion by masking 75% tubelets or fully dropping a selected plane (100% masking), which corresponds to the inter-plane masking case used in pretraining. The model reconstructs the missing content from the remaining planes via the shared latent volume, and the recovered boundaries in unobserved views suggest genuine global 3D reasoning rather than intra-plane 2D inpainting. This capability suggests potential for robust inference under sparse acquisitions or missing planes, e.g., latent interpolation with diffusion models [2].

Multi-plane Segmentation. Table 1 and Fig. 3 report multi-plane cine CMR segmentation results. We consider two supervision protocols. The standard protocol uses view-specific label sets for SAX and LAX (followed by UNETR+

Table 1: Segmentation Dice scores of nnUNet, UNETR+, WholeNet, and our models. nnUNet is trained per plane type (single 2D+t input; SAX/LAX separately), whereas UNETR+, WholeNet, and ours use all available sparse planes as input. Best is in bold and second best is underlined.

Phenotype	nnUNet	UNETR+	WholeNet	Ours	Ours†
LVBP	$\underline{0.95_{\pm 0.02}}$	$0.86_{\pm 0.02}$	$0.93_{\pm 0.02}$	$\mathbf{0.96}_{\pm 0.02}$	$\mathbf{0.97}_{\pm 0.01}$
LVMYO	$\mathbf{0.94}_{\pm 0.01}$	$0.83_{\pm 0.01}$	$0.86_{\pm 0.04}$	$\underline{0.92_{\pm 0.02}}$	$\mathbf{0.93}_{\pm 0.02}$
RVBP	$\mathbf{0.94}_{\pm 0.03}$	$0.88_{\pm 0.03}$	$0.89_{\pm 0.04}$	$\mathbf{0.94}_{\pm 0.03}$	$\mathbf{0.96}_{\pm 0.02}$
LABP	$\underline{0.95_{\pm 0.02}}$	$0.85_{\pm 0.02}$	$0.91_{\pm 0.03}$	$\mathbf{0.96}_{\pm 0.02}$	$\mathbf{0.97}_{\pm 0.02}$
RABP	$\underline{0.94_{\pm 0.02}}$	$0.91_{\pm 0.02}$	$0.92_{\pm 0.02}$	$\mathbf{0.96}_{\pm 0.04}$	$\mathbf{0.98}_{\pm 0.02}$

Table 2: Ablation study on data scaling paradigms. We compare validation Dice scores under varying sizes of unlabeled pretraining and labeled segmentation datasets. Best is in bold and second best is underlined.

Setting	Pretrain		Seg		Dice Score				
	1k	4k	1k	4k	LVBP	LVMYO	RVBP	LABP	RABP
C1	✓		✓		0.9666	0.9135	0.9452	0.9589	0.9700
C2	✓			✓	0.9705	<u>0.9212</u>	0.9482	0.9654	0.9750
C3		✓	✓		<u>0.9718</u>	0.9215	<u>0.9494</u>	<u>0.9675</u>	<u>0.9759</u>
C4		✓		✓	**0.9840**	**0.9555**	**0.9668**	**0.9786**	**0.9835**

and WholeNet), while Ours†[6] adopts a unified cross-plane label space where the same anatomy shares a consistent class definition on both SAX and LAX despite view-dependent 2D appearances. Under the standard protocol, Ours consistently ranks first or second across regions and yields clear gains over nnUNet, UNETR+, and WholeNet, corroborating that geometry-aware volume pretraining transfers effectively to segmentation. Under the unified protocol, Ours† further improves even though the supervision is more challenging, indicating that explicit plane-to-volume modelling learns consistent 3D semantics that reduce view-induced ambiguity and strengthen cross-plane supervision.

Scalability and Label Efficiency. We study how segmentation result varies with the size of the unlabeled pretraining cohort and the labeled segmentation set. Table 2 shows that larger pretraining consistently improves Dice score, with the clearest gains under limited labels, indicating that geometry-aware pretraining reduces annotation dependence by transferring robust 3D priors. With more labeled data, the best performance is achieved by the largest configuration.

[6] Ours† uses a unified label space across planes.

5 Conclusion

In this study, we presented a geometry-aware self-supervised pretraining framework for learning whole-heart representations from sparse multi-plane cine CMR. By injecting patient-space coordinates, encoding planes with a shared encoder, and enforcing explicit plane-to-volume fusion with coordinate-constrained reprojection, the learned latent volume captures coherent 3D structure and enables reliable cross-plane completion. Experiments on UK Biobank demonstrate strong transfer to downstream multi-plane segmentation and robustness to sparse-plane inputs.

Acknowledgments. L. Hu was supported by the China Scholarship Council (CSC). This work was funded in part by Imperial College London I-X and the Imperial College London Seeds for Success Fund. This research has been conducted using the UK Biobank Resource under Application Number 100203. G. Yang was supported in part by the ERC IMI (101005122), the H2020 (952172), the MRC (MC/PC/21013), the Royal Society (IEC/NSFC/211235), the NVIDIA Academic Hardware Grant Program, the SABER project supported by Boehringer Ingelheim Ltd, the NIHR Imperial Biomedical Research Centre (RDA01), the Wellcome Leap Dynamic Resilience program (co-funded by Temasek Trust), UKRI guarantee funding for Horizon Europe MSCA Postdoctoral Fellowships (EP/Z002206/1), a UKRI MRC Research Grant, TFS Research Grants (MR/U506710/1), the Swiss National Science Foundation (Grant No. 220785), and the UKRI Future Leaders Fellowship (MR/V023799/1, UKRI2738).

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

References

1. Bai, W., Suzuki, H., Huang, J., Francis, C., Wang, S., Tarroni, G., Guitton, F., Aung, N., Fung, K., Petersen, S.E., et al.: A population-based phenome-wide association study of cardiac and aortic structure and function. Nature Medicine **26**(10), 1654–1662 (2020). https://doi.org/10.1038/s41591-020-1009-y
2. Bubeck, N., Shit, S., Chen, C., Zhao, C., Guo, P., Yang, D., Zitzlsberger, G., Xu, D., Kainz, B., Rueckert, D., et al.: Latent interpolation learning using diffusion models for cardiac volume reconstruction (2025). https://doi.org/10.48550/ARXIV.2508.13826
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (2020)
4. Chen, Y., Yang, J., Mercadier, D.S., Le, H., Schwitter, J., Fua, P.: End-to-end 4D heart mesh recovery across full-stack and sparse cardiac MRI (2025). https://doi.org/10.48550/ARXIV.2509.12090
5. Ding, Z., Hu, Y., Li, Z., Zhang, H., Wu, F., Xiang, Y., Li, T., Liu, Z., Chu, X., Huang, Z.: Cross-modality cardiac insight transfer: A contrastive learning approach to enrich ECG with CMR features. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. p. 109–119. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-72384-1_11

6. Feichtenhofer, C., Fan, h., Li, Y., He, K.: Masked autoencoders as spatiotemporal learners. In: Advances in Neural Information Processing Systems. vol. 35, pp. 35946–35958. Curran Associates, Inc. (2022)
7. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 574–584 (2022)
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (2022)
9. Isensee, F., Jaeger, P.F., Kohl, S.A.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 (2020). https://doi.org/10.1038/s41592-020-01008-z
10. Jafari, M., Shoeibi, A., Khodatars, M., Ghassemi, N., Moridian, P., Alizadehsani, R., Khosravi, A., Ling, S.H., Delfan, N., Zhang, Y.D., et al.: Automated diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging using deep learning models: A review. Computers in Biology and Medicine **160**, 106998 (2023). https://doi.org/10.1016/j.compbiomed.2023.106998
11. Li, Y., Yu, Z., Choy, C., Xiao, C., Alvarez, J.M., Fidler, S., Feng, C., Anandkumar, A.: Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9087–9098 (2023)
12. Pennell, D.J.: Cardiovascular magnetic resonance. Circulation **121**(5), 692–705 (2010). https://doi.org/10.1161/circulationaha.108.811547
13. Petersen, S.E., Matthews, P.M., Francis, J.M., Robson, M.D., Zemrak, F., Boubertakh, R., Young, A.A., Hudson, S., Weale, P., Garratt, S., et al.: UK biobank's cardiovascular magnetic resonance protocol. Journal of Cardiovascular Magnetic Resonance **18**(1), 8 (2016). https://doi.org/10.1186/s12968-016-0227-4
14. Selivanov, A., Müller, P., Turgut, O., Stolt-Ansó, N., Rueckert, D.: Global and local contrastive learning for joint representations from cardiac MRI and ECG. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. p. 217–227. Springer Nature Switzerland (2025). https://doi.org/10.1007/978-3-032-04927-8_21
15. Shah, K., Crandall, R., Xu, J., Zhou, P., George, M., Bansal, M., Chellappa, R.: Mv2mae: Multi-view video masked autoencoders (2024). https://doi.org/10.48550/ARXIV.2401.15900
16. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Advances in Neural Information Processing Systems. vol. 35, pp. 10078–10093. Curran Associates, Inc. (2022)
17. Turgut, O., Müller, P., Hager, P., Shit, S., Starck, S., Menten, M.J., Martens, E., Rueckert, D.: Unlocking the diagnostic potential of electrocardiograms through information transfer from cardiac magnetic resonance imaging. Medical Image Analysis **101**, 103451 (2025). https://doi.org/10.1016/j.media.2024.103451
18. Ugurlu, D., Qian, S., Fairweather, E., Mauger, C., Ruijsink, B., Toso, L.D., Deng, Y., Strocchi, M., Razavi, R., Young, A., et al.: Cardiac digital twins at scale from MRI: Open tools and representative models from˜ 55000 UK biobank participants. PLOS One **20**(7), e0327158 (2025). https://doi.org/10.1371/journal.pone.0327158

19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
20. Wang, Y.R., Yang, K., Wen, Y., Wang, P., Hu, Y., Lai, Y., Wang, Y., Zhao, K., Tang, S., Zhang, A., et al.: Screening and diagnosis of cardiovascular disease using artificial intelligence-enabled cardiac magnetic resonance imaging. Nature Medicine **30**(5), 1471–1480 (2024). https://doi.org/10.1038/s41591-024-02971-2
21. Wang, Z., Huang, M., Shi, Z., Hu, H., Lan, L., Zhang, H., Li, Y., Hu, X., Lu, Q., Zhu, Z., et al.: Enabling ultra-fast cardiovascular imaging across heterogeneous clinical environments with a generalist foundation model and multimodal database (2025). https://doi.org/10.48550/ARXIV.2512.21652
22. Yan, S., Xiong, X., Arnab, A., Lu, Z., Zhang, M., Sun, C., Schmid, C.: Multiview transformers for video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3333–3343 (2022)
23. Yu, Z., Zhang, R., Ying, J., Yu, J., Hu, X., Luo, L., Cao, S.Y., Shen, H.L.: Context and geometry aware voxel transformer for semantic scene completion. In: Advances in Neural Information Processing Systems. vol. 37, pp. 1531–1555. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-0049
24. Zhang, Y., Chen, C., Shit, S., Starck, S., Rueckert, D., Pan, J.: Whole heart 3d+t representation learning through sparse 2d cardiac mr images. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. p. 359–369. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-72378-0_34