Anatomy-Prior Guided Decoupled Expert Networks for Multi-View Cardiac MRI Segmentation and Biomarker Estimation

Wenpei Liu[1], Zhihan Zhang[1], Zehao Yang[1], Yujia Song[1], Chenxi Li[1], Xinyang Liu[1], Yulin Sun[1], and Runnan He[1,2,3⋆], Xiuyun Liu[1,2,3,4*], Dong Ming[1,2,3]

[1] Medical School, Tianjin University, Tianjin, China
[2] State Key Laboratory of Advanced Medical Materials and Devices, Tianjin University, Tianjin, China
[3] Haihe Laboratory of Brain-Computer Interaction and Human-Machine Integration, Tianjin, China
[4] School of Pharmaceutical Science and Technology, Tianjin University,Tianjin, China
runnanhe@tju.edu.cn, xiuyun_liu@tju.edu.cn

Abstract. Automated multi-view cardiac MRI segmentation and downstream biomarker estimation are challenged by sequence heterogeneity, sparse pathological targets, and negative transfer in shared multi-head models. We propose an anatomy-prior guided system of view-specific experts, using 3D Swin UNETR for resized volumetric Cine inputs and 2D TransUNet with one-vs-rest supervision for LGE MRI. An orientation-preserving augmentation policy disables vertical flipping on long-axis views. On the official MICCAI 2026 CMR-MULTI validation set, the Cine ensemble achieves Dice scores of 85.31% (SAX), 82.39% (2CH), and 84.68% (4CH), while the LGE ensemble achieves 66.32%, 63.31%, 76.56%, and 83.71% on SAX, 2CH, 4CH, and RAS, respectively. A controlled shared-encoder comparison supports the contribution of view-specific optimization. Validation-set analysis yields an LVEF MAE of 6.04% and a myocardial-mass MAE of 5.52 g; however, systematic underestimation of absolute Cine volumes remains a limitation.The code will be available at https://github.com/LwpZoe/CMR.

Keywords: Cardiac MRI · Multi-View Segmentation · TransUNet · Anatomical Prior · Biomarker Prediction

1 Introduction

Cardiac magnetic resonance (CMR) imaging is the clinical gold standard for assessing cardiac function and pathology, with LVEF and myocardial mass serving as important biomarkers [9]. Automated segmentation across multi-view Cine and LGE sequences remains difficult because of sequence heterogeneity, anatomical variation, and multi-center domain shifts [11].

⋆ Corresponding authors: Runnan He and Xiuyun Liu.

Existing frameworks generally follow two paradigms: monolithic shared-encoder multi-head architectures and standard 2D slice-by-slice processing. Shared-encoder models often trigger gradient conflicts, where heterogeneous view features (e.g., SAX rings vs. long-axis cavities) cause negative transfer, destabilizing training [5–7]. Conversely, standard 2D approaches often ignore anatomical spatial priors, leading to physically impossible predictions like upside-down hearts in long-axis views [14, 8, 10]. Furthermore, 3D Cine analysis often employs naive downsampling that collapses spatiotemporal interleaving, resulting in boundary blurring [2, 13].

To mitigate these limitations, we propose a decoupled expert framework in which heterogeneous views are optimized independently rather than through a shared encoder. For Cine MRI (Task 1), a 3D Swin UNETR [2] models volumetric dependencies in the resized input. For LGE MRI (Task 2), a 2D TransUNet [1] combines slice-wise normalization, one-vs-rest supervision, and an orientation-preserving augmentation policy. Five-fold ensembles and physical-space reconstruction support segmentation and downstream biomarker estimation. Our contributions are: (1) a controlled comparison of shared and view-specific TransUNet models; (2) an ablation of the long-axis orientation prior; and (3) validation-set analyses of class-wise volume error, biomarker agreement, computational cost, and representative failure cases.

2 Methodology

2.1 Overall Dual-Track Framework

Let $\mathcal{D}_{m,v} = \{(x_i, y_i)\}_{i=1}^{N_{m,v}}$ denote the data for modality $m \in \{\text{Cine}, \text{LGE}\}$ and anatomical view v. Instead of minimizing a joint shared-encoder objective over heterogeneous views, we independently optimize parameters $\theta_{m,v}$ for each expert. This avoids direct gradient interaction between view-specific objectives and is evaluated against a controlled shared-encoder model. As depicted in Fig. 1, the system comprises (1) a 3D Swin UNETR ensemble for Cine MRI and (2) a 2D TransUNet ensemble for LGE segmentation and mass estimation.

The backbones were selected according to the dimensional characteristics of the modalities rather than an assumption of universal architectural superiority. Swin UNETR provides hierarchical volumetric context with memory-efficient shifted-window attention for Cine MRI, whereas 2D TransUNet supports slice-wise processing of anisotropic LGE volumes while combining local CNN features with global Transformer context. The official Cine baseline is a shared-encoder 3D CNN with view-specific decoder heads. Its ResUNet++-style design uses residual 3D convolutional blocks, squeeze-and-excitation modules, an ASPP3D bottleneck, and attention-gated skip connections. Alternatives such as full-resolution 3D nnU-Net remain valid choices; the present comparison uses this CNN baseline without claiming universal superiority over all CNN or nnU-Net configurations.

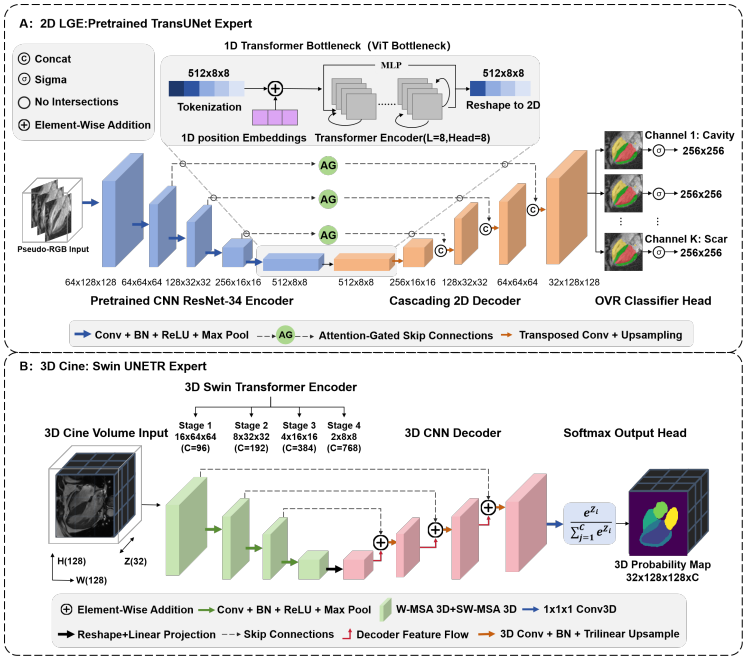

Fig. 1. Architectures of the view-specific experts. (A) The 2D TransUNet expert for LGE uses a ResNet-34 encoder, an eight-layer Transformer bottleneck with eight attention heads, and a cascaded decoder with attention gates and one-vs-rest outputs. (B) The 3D Swin UNETR expert for Cine uses hierarchical shifted-window attention to model volumetric dependencies and a symmetric 3D CNN decoder with multi-scale skip connections.

2.2 Task 1: 3D Swin UNETR Ensemble for Cine MRI

Spatiotemporal Downsampling: To manage GPU memory while preserving temporal continuity, input volumes $I_{\text{Cine}} \in \mathbb{R}^{H\times W\times Z}$ are normalized via Z-score and scaled to a fixed grid $\mathcal{X}_{\text{Cine}} \in \mathbb{R}^{D'\times H'\times W'}$ $(32 \times 128 \times 128)$ using trilinear interpolation $\mathcal{T}_{\text{tri}}$:

$$\mathcal{X}_{\text{Cine}} = \mathcal{T}_{\text{tri}}\left(\mathcal{N}_{\text{Z}}\left(I_{\text{Cine}}\right), (D', H', W')\right). \tag{1}$$

Hierarchical 3D Swin Encoder: The volume $\mathcal{X}'_{\text{Cine}}$ is partitioned into $2\times2\times2$ patches and processed by a 3D Swin Transformer encoder [2]. Shifted window self-attention (W-MSA/SW-MSA) is computed at layer l to model long-range dependencies:

$$\hat{\mathbf{z}}^l = \text{W-MSA}\left(\text{LN}\left(\mathbf{z}^{l-1}\right)\right) + \mathbf{z}^{l-1}, \quad \mathbf{z}^l = \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^l\right)\right) + \hat{\mathbf{z}}^l. \tag{2}$$

Multi-scale features are fused via a 3D CNN decoder to recover boundaries. We optimize the network using a compound Dice-CE loss: $\mathcal{L}_{\text{Cine}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Dice}}$.

2.3 Task 2: 2D TransUNet Ensemble for LGE MRI

Anatomy-Aware Strategy: To preserve hyper-enhancement contrast, LGE slices are normalized locally and adjacent slices are stacked as three-channel pseudo-RGB inputs. We disable vertical flips on 2CH, 4CH, and RAS views as an orientation-preserving prior intended to avoid superior-inferior reversals. Table 1 shows that its quantitative effect is view-dependent, with the clearest overlap and boundary improvement observed for 4CH.
Architecture and OVR Loss: We use a TransUNet [1] with a ResNet-34 encoder [3] and an eight-layer Transformer bottleneck. Independent one-vs-rest (OVR) heads represent the K foreground categories. The optimization objective is

$$\mathcal{L}_{\text{LGE}} = 0.5\,\mathcal{L}_{\text{BCE-F}} + 0.5\,\mathcal{L}_{\text{Dice-OVR}}, \tag{3}$$

where $\mathcal{L}_{\text{BCE-F}}$ is binary cross-entropy modulated by a focal factor with weight $\alpha = 0.4$ and exponent $\gamma = 2.0$, and $\mathcal{L}_{\text{Dice-OVR}}$ is the mean binary SoftDice loss over foreground heads. This formulation retains BCE supervision while increasing the contribution of difficult pixels and sparse pathological structures.

2.4 Ensemble Inference and 3D Reconstruction

Final predictions are generated via 5-fold soft-voting: $\bar{P}_c = \frac{1}{M}\sum_{m=1}^{M}\sigma(\hat{y}_c^{(m)})$. To bridge the 2D-3D gap, we perform in-situ reconstruction by extracting original NIfTI affine matrices $\mathbf{A}$ and spacing headers $\mathbf{S}_{xyz}$ from patient metadata:

$$V_{\text{pred}} = \mathcal{T}_{\text{near}}\left(\mathbf{P}_{\text{2D}}, (H_{\text{orig}}, W_{\text{orig}}, Z_{\text{orig}})\right) \otimes (\mathbf{A}, \mathbf{S}_{xyz})\,. \tag{4}$$

Clinical biomarkers are then derived from the reconstructed volumes. LVEF is computed as a ratio of stroke volume to end-diastolic volume (EDV):

$$LVEF = \frac{EDV - ESV}{EDV} \times 100\%, \tag{5}$$

where EDV and ESV are the maximum and minimum ventricular volumes across the cardiac cycle. Because resizing introduces absolute scale bias, validation-set agreement is reported in Supplementary Table S7.

3 Experiments

3.1 Experimental Setup

Experiments were conducted on the MICCAI 2026 Universal Multi-Sequence, Multi-Center, and Multi-View CMR Segmentation Challenge (CMRSeg) dataset [11]. The dataset covers Cine MRI (Task 1) and late gadolinium enhancement (LGE) MRI (Task 2) across short-axis (SAX), two-chamber (2CH), and four-chamber (4CH) views. The LGE track additionally includes right atrial segmentation using the RAS dataset [12]. We followed the official training, validation, and test partitions.

Table 1. Ablation of vertical flipping on long-axis LGE segmentation. Results are mean ± standard deviation over five folds. HD95 is evaluated on the resized grid using the same unit-spacing protocol for both settings; lower is better.

View	Vertical Flip	Dice ↑	HD95 ↓
2CH	Enabled	$\mathbf{0.6288 \pm 0.0368}$	117.98 ± 31.09
	Disabled	0.6007 ± 0.0384	$\mathbf{116.24 \pm 45.59}$
4CH	Enabled	0.6558 ± 0.0377	103.63 ± 28.47
	Disabled	$\mathbf{0.6830 \pm 0.0408}$	$\mathbf{91.61 \pm 28.60}$

Experiments were conducted using two NVIDIA RTX 5090 GPUs with 32 GB memory under CUDA 12.8. Both tasks were optimized using Adam with an initial learning rate of 10^{-4} and a weight decay of 10^{-5} for up to 300 epochs. For LGE segmentation, we used a batch size of 4, a cosine annealing learning-rate scheduler, and early stopping with a patience of 50 epochs. For Cine segmentation, we used a batch size of 1 and a ReduceLROnPlateau scheduler. Cine volumes were resized to $32 \times 128 \times 128$ to satisfy the 2^5 divisibility requirement of Swin UNETR.

3.2 Training Stability and Internal Validation

Five-fold cross-validation was performed within the official training partition. Mean Dice was 0.6464, 0.6007, 0.6830, and 0.8619 for the LGE SAX, 2CH, 4CH, and RAS experts, respectively, and 0.8422, 0.8200, and 0.8218 for the Cine SAX, 2CH, and 4CH experts. Complete fold-wise results are provided in Supplementary Tables S1-S2.

3.3 Ablation of the Anatomy-Aware Augmentation

To isolate the contribution of the orientation-preserving augmentation strategy, we retrained the 2CH and 4CH experts while enabling random vertical flips, keeping the architecture, loss, data partitions, and optimization settings unchanged. Table 1 reports the fold-matched results over five folds. Disabling vertical flips improved 4CH Dice from 0.6558 to 0.6830 and reduced HD95 from 103.63 to 91.61. For 2CH, vertical flipping increased Dice from 0.6007 to 0.6288, whereas HD95 remained similar. Averaged across both long-axis views, disabling vertical flips produced nearly unchanged Dice (0.6419 vs. 0.6423) but reduced HD95 from 110.81 to 103.93. These findings indicate that the anatomical orientation constraint primarily improves boundary stability and has a view-dependent effect, with the clearest benefit observed in 4CH.

3.4 Controlled Shared-Encoder Baseline Analysis

To distinguish view-specific optimization from a simple backbone replacement, we compare the shared-encoder TransUNet and CNN (ResUNet++) baseline with the view-specific experts in Table 2. The shared TransUNet improves 4CH but performs less consistently on SAX and 2CH, whereas the experts achieve the highest overall mean Dice. The comparison is not strictly parameter matched: the shared TransUNet contains 50.72M parameters, while one active expert contains 25.13M; it is therefore an end-to-end system comparison rather than an isolated backbone ablation.

Table 2. Controlled comparison between a CNN baseline, a shared TransUNet baseline, and the proposed decoupled TransUNet experts on LGE MRI segmentation.

Method	SAX	2CH	4CH	RAS	**Mean**
CNN (ResUNet++)	0.5595	0.5748	0.5996	0.8634	0.6493
Shared TransUNet	0.4529	0.4815	**0.7514**	0.8068	0.6232
Decoupled Experts	**0.6464**	**0.6007**	0.6830	**0.8619**	**0.6980**

3.5 Cross-Center Generalization Performance

We compare the proposed experts with the multi-head baseline on the independent validation set in Table 3. LGE performance improves most clearly in 4CH (Dice +16.60 percentage points and HD95 −99.5 under the common evaluation protocol), consistent with reduced negative transfer under view-specific optimization. For Cine MRI, the baseline retains higher Dice, whereas the expert ensemble yields lower HD95 across all views. This establishes an overlap-boundary trade-off rather than uniform superiority.

Table 3. Performance comparison between the official multi-head baseline and our proposed expert ensemble on VAL set (Mean Dice and HD95).

Modality	Method	SAX		2CH		4CH	
		Dice	HD95	Dice	HD95	Dice	HD95
LGE (2D)	CNN (ResUNet++)	0.5595	234.25	0.5748	169.80	0.5996	163.55
	Ours	**0.6632**	**133.66**	**0.6331**	**128.71**	**0.7656**	**64.05**
Cine (3D)	CNN (ResUNet++)	**0.8853**	2.54	**0.8521**	4.10	**0.8811**	2.69
	Ours	0.8531	**2.36**	0.8239	**2.24**	0.8468	**2.39**

Class-wise analysis identifies LGE 2CH scar segmentation as the main remaining pathology limitation (Dice 0.4012). Complete class-wise metrics, Cine

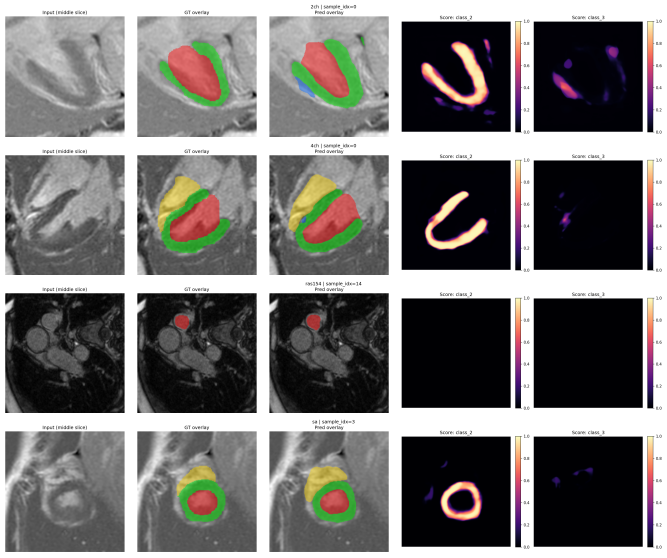

Fig. 2. Representative LGE validation results. Each row shows an input slice and the corresponding reference, prediction, and class-probability overlays. The examples illustrate both successful boundary delineation and residual myocardium-scar ambiguity.

results, and qualitative examples are provided in Supplementary Tables S4-S5 and Fig. S1.

3.6 External Inference and Operational Feasibility

The Docker pipeline processed all 30 Cine and 13 LGE official test cases without runtime failure [11]. Predicted LVEF ranged from 27.0% to 59.8%, and myocardial mass ranged from 34.5 to 170.9 g. Because labels were unavailable, these ranges demonstrate operational feasibility rather than clinical accuracy. Per-case predictions are listed in Supplementary Tables S8-S9.

4 Discussion

The experiments reveal both benefits and limitations of view-specific experts. In Cine MRI, the proposed model achieves lower HD95 than the official baseline across all views, but its Dice scores are lower and resized-space volume estimates show systematic negative bias. The method therefore exhibits an overlap–boundary trade-off rather than uniform superiority. The validation-set biomarker analysis further clarifies this limitation. Although Cine EDV and ESV are undere stimated, LVEF remains relatively stable because it is computed as a ratio between EDV and ESV (Eq. 5), achieving an MAE of 6.04% and Pearson correlation of 0.859 on the validation set. Thus, the current pipeline is more reliable

for relative functional assessment than for absolute volumetric quantification. Future research will focus on resolving absolute scale shrinkage through motion-aware spatiotemporal decoupling, as suggested in Mesh4D[15].

For LGE MRI, the controlled shared-encoder comparison suggests that the gain is not solely attributable to the TransUNet backbone. The augmentation ablation further shows that orientation preservation is view dependent: it improves both Dice and HD95 for 4CH but not 2CH Dice. Meanwhile, moderate 2CH scar Dice and visible myocardium–scar confusion show that fine-grained pathology quantification remains difficult. The 20-model ensemble also increases offline training and inference cost, despite the lower active parameter count of one expert.

5 Conclusion

We presented an anatomy-prior guided system of view-specific experts for multi-view CMR segmentation and biomarker estimation. Controlled comparison with a shared TransUNet supports the value of view-specific optimization, while the augmentation ablation demonstrates a view-dependent orientation effect with the clearest benefit in 4CH. Validation-set analysis identifies lower Cine HD95 but also systematic absolute-volume compression, whereas LVEF and LGE myocardial mass show stronger agreement with reference masks. The containerized pipeline processed the official test set successfully, but further physical-volume calibration and external clinical validation are required before deployment.

Acknowledgements. This work was supported by the Scientific Research Innovation Capability Support Project for Young Faculty (ZYGXQNJSKYCXNLZCXM-H15); the National Natural Science Foundation of China (Nos. 82472098 and 32300704); the Tianjin Natural Science Foundation Outstanding Youth Project (No. 24JCJQJC00250) and General Program (No. 25JCLMJC00030); the Major Science and Technology Special Project (No. 25ZXCKQY00100); the Shandong Provincial Natural Science Foundation Major Basic Research Project (No. ZR2025ZD13); the National Key Research and Development Program of China (No. 2025YFE0214400); the Autonomous Project of Haihe Laboratory of Brain-Computer Interaction and Human-Machine Integration (No. 24HHNJSS00003, No. 24HHNJSS00005, No. 24HHNJSS00010, and 2025E7-0121); and the Non-profit Central Research Institute Fund of the Chinese Academy of Medical Sciences (No. 2024-JKCS-16).

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

References

1. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv preprint arXiv:2102.04306 (2021)

2. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. In: International MICCAI Brainlesion Workshop. pp. 272-284. Springer, Cham (2021)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770-778 (2016)
4. Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D.: ResUNet++: An Advanced Architecture for Medical Image Segmentation. In: 2019 IEEE International Symposium on Multimedia (ISM). pp. 225-232. IEEE (2019)
5. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient Surgery for Multi-Task Learning. Advances in Neural Information Processing Systems (NeurIPS) **33**, 5824-5836 (2020)
6. Ma, J., Han, L., Jia, D., Xie, H., Tan, T., Tong, H.H., Lee, A.W., Soong, S.I., Sun, Y.: Grad-MTSeg: Mitigating Multi-Task Gradient Conflicts via Hierarchical Gradient Optimization for NPC Radiotherapy Delineation. In: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3926-3929. IEEE (2025)
7. Yang, J., Pang, X., Lin, C., Tan, T.: MOTDNet: Multi Organ Task Decoupling Network for Cell Segmentation. Medical Image Analysis **104**, 104045 (2026)
8. Gao, Z., Zhou, F., Zhao, H., Chen, W., Luo, R., Wang, X., Qin, H., Wu, Z., Meng, Y., Zhao, Y., et al.: HadBalance: A Plug-and-Play Unified Global Geometric Prior Framework for Generalizable Biomedical Segmentation. arXiv preprint arXiv:2606.15976 (2026)
9. Xu, R., Jiang, S., Zhai, Y., Chen, Y.: CorSeg-CineSAX: An Open-Source Deep Learning Framework for Fully Automatic Segmentation of Short-Axis Cine Cardiac MRI Across Multiple Cardiac Diseases. medRxiv preprint (2026). https://doi.org/10.1101/2026.04.28.26354452
10. Zhang, J., Bergere, B., Bollache, E., Leite, J., Laredo, M., Redheuil, A., Kachenoura, N.: Progressive Learning with Anatomical Priors for Reliable Left Atrial Scar Segmentation from Late Gadolinium Enhancement MRI. arXiv preprint arXiv:2603.26186 (2026)
11. Qu, T., Su, Z., Zhang, H., et al.: Universal Multi-Sequence, Multi-Center and Multi-View CMR Segmentation Challenge.MICCAI 2026, Zenodo (2026). `https://doi.org/10.5281/zenodo.19728181`
12. Zhu, J., Bai, J., Zhou, Z., et al.: RAS Dataset: A 3D Cardiac LGE-MRI Dataset for Segmentation of the Right Atrial Cavity. Scientific Data **11**(1), 401 (2024)
13. Płotka, S., Chrabaszcz, M., Biecek, P.: Swin SMT: Global Sequential Modeling for Enhancing 3D Medical Image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 689-698. Springer, Cham (2024)
14. El Jurdi, R., Petitjean, C., Honeine, P., Cheplygina, V., Abdallah, F.: High-Level Prior-Based Loss Functions for Medical Image Segmentation: A Survey. Computer Vision and Image Understanding **210**, 103248 (2021)
15. Qiao, M., Zheng, J., Zhang, W., Ma, Q., Li, L., Kainz, B., O'Regan, D.P., Matthews, P.M., Niederer, S., Bai, W.: Mesh4D: A Motion-Aware Multi-View Variational Autoencoder for 3D+t Mesh Reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 343-353. Springer, Cham (2025)