MaskSAM-PBPR: A Prior-Guided Bidirectional Propagation Refiner for Multi-Sequence CMR Myocardial Pathology Segmentation

Qinghe Zhang, Yanjun Peng✉, Xiaoning Zhang, and Fuzhuang Qiu

Shandong University of Science and Technology, 579 Qianwan Gang Road, 266590 Huangdao, Qingdao, Shandong, China
{zhangqinghe,yjpeng,zhangxiaoning,202123030301}@sdust.edu.cn

Abstract. Multi-sequence cardiac magnetic resonance (CMR) provides complementary tissue contrasts. However, myocardial edema and scar remain difficult to segment because they are small, poorly delineated, and may be discontinuous across slices. Cross-center intensity shifts and imperfect inter-sequence alignment further complicate the joint use of late gadolinium enhancement (LGE), T2-weighted, and balanced steady-state free precession (bSSFP) images. Query-based segmenters can localize irregular lesions, but slice-wise query predictions do not explicitly enforce volumetric coherence. To address this limitation, we propose MaskSAM-PBPR, a two-stage framework that separates query-based localization from prior-guided volumetric refinement. In Stage 1, the Enhanced MaskSAM Prior Generator uses MedSAM initialization, prompt denoising, and Fourier-domain augmentation to produce class-specific soft priors. In Stage 2, the Prior-Guided Bidirectional Propagation Refiner (PBPR) combines Anchor-Guided Class-Wise Propagation with 3D Residual Refinement. This stage converts slice-wise soft priors into volumetric predictions by propagating class-specific evidence across slices and correcting residual local errors at full resolution. On the CARE 2026 validation set, MaskSAM-PBPR achieved scar and edema Dice scores of 0.7200 and 0.7322, with Hausdorff distance (HD) values of 12.5062 and 19.7994 mm.

Keywords: Multi-sequence CMR · Myocardial pathology · MaskSAM · Bidirectional propagation · Residual refinement

1 Introduction

Cardiac magnetic resonance (CMR) provides complementary views of myocardial anatomy and tissue injury. Late gadolinium enhancement (LGE) depicts scar and T2-weighted CMR highlights edema [9,15], whereas balanced steady-state free precession (bSSFP) delineates cardiac anatomy [11,4]. Accurate automatic segmentation can reduce annotation burden and improve the reproducibility of myocardial pathology quantification [9].

The MyoPS setting is complicated by cross-center variation in pathology appearance arising from differences in scanners and acquisition protocols [4,14].

Edema and scar occupy only a small volume fraction and often have weak or ambiguous boundaries [9,5]. Their anatomical interdependence further complicates prediction because they cannot be treated as fully independent semantic categories [5]. Together, these factors challenge both voxel-wise prediction and query-based segmentation [13]. In query-based models, slice-wise set prediction also lacks an explicit volumetric constraint, which can leave neighboring predictions inconsistent.

Existing MyoPS methods principally improve input harmonization and multi-sequence representation. MyoPS-Net supports flexible combinations of CMR sequences [11], whereas explicit sequence alignment reduces residual spatial mismatch between modalities [4]. Recent CARE 2025 methods further emphasize cross-center robustness and boundary recovery. EHU-Mamba2 combines dynamic histogram matching, attention-based feature selection, and adaptive upsampling to address domain shifts and ambiguous lesion boundaries [14]. These advances target modality availability, registration, or local feature recovery, but do not address the specific problem of propagating query-derived priors across slices.

DETR-style models predict sets of masks and classes [1,2]. SAM established promptable segmentation, and MedSAM adapted this paradigm for medical images [7,10]. MaskSAM extends mask classification to volumetric medical images [13]. Unlike dense voxel classifiers, query-based segmentation provides an explicit set of candidate regions and class scores. However, the identity and confidence of these candidates can vary between neighboring slices. In our experiments, a single-stage MaskSAM baseline consequently produced fragmented lesion responses along the slice axis.

To address this gap, we divide MyoPS segmentation into two complementary stages. Stage 1 localizes candidate regions and produces class-specific soft priors, whereas Stage 2 models cross-slice context and refines local boundaries. Our contributions follow this two-stage design. First, in Stage 1, the Enhanced MaskSAM Prior Generator combines query-wise decoder training, denoising with perturbed prompts, and Fourier-domain augmentation to produce class-specific soft priors. Second, in Stage 2, PBPR jointly applies Anchor-Guided Class-Wise Propagation and 3D Residual Refinement. The former integrates a hybrid slice encoder, prior-derived anchor initialization, and independent bidirectional ConvGRUs to model through-plane context, whereas the latter corrects residual errors and refines local boundaries at full resolution.

2 Method

2.1 Pipeline Overview

Given a three-sequence CMR volume $X \in \mathbb{R}^{3\times D\times H\times W}$ containing LGE, T2-weighted, and bSSFP images, the goal is to predict $Y \in \{0,1,2,3\}^{D\times H\times W}$ for background, myocardium, edema, and scar.

As shown in Fig. 1, MaskSAM-PBPR comprises prior generation, leakage-free training-input construction, and PBPR-based volumetric refinement. Stage 1

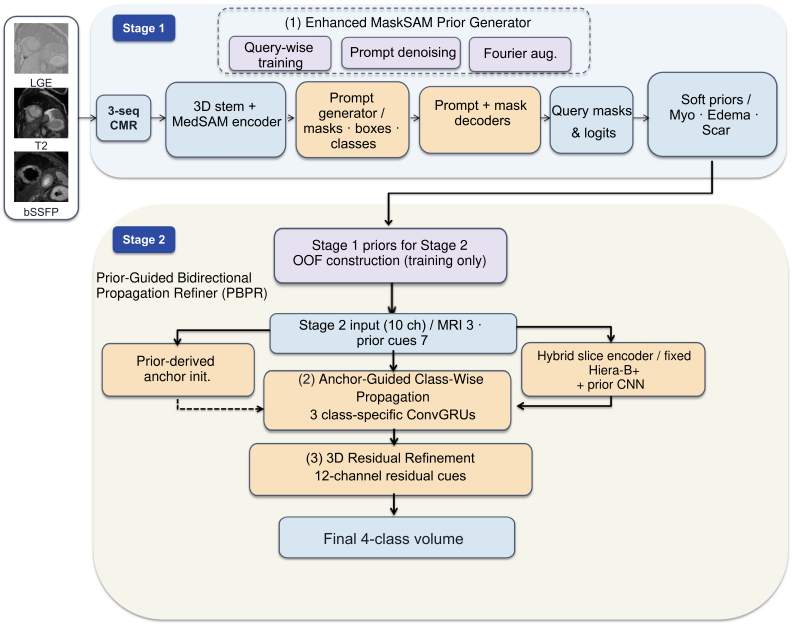

Fig. 1. The pipeline of MaskSAM-PBPR. Stage 1 is the Enhanced MaskSAM Prior Generator. OOF predictions are used only to construct leakage-free Stage 2 training inputs; at inference, the same representation is built from the current Stage 1 priors. PBPR combines Anchor-Guided Class-Wise Propagation with 3D Residual Refinement.

predicts class-specific soft priors from the three CMR sequences. During training, out-of-fold (OOF) predictions are combined with the source images to form a 10-channel refinement tensor. At inference, the same representation is constructed from the corresponding Stage 1 predictions. Stage 2 then applies Anchor-Guided Class-Wise Propagation and 3D Residual Refinement to improve through-plane continuity and local boundaries.

2.2 Stage 1: Enhanced MaskSAM Prior Generator

Stage 1 provides the coarse segmentation and soft priors. A shallow 3D convolutional stem maps the three-sequence input to a SAM-compatible representation, and a MedSAM-initialized image encoder extracts multi-scale features [10]. A query decoder predicts auxiliary masks, boxes, and classifier tokens. The prompt encoder and mask decoder then produce $Q = 8$ query masks and corresponding class logits. Following MaskSAM [13], one binary target is constructed for each semantic label on each slice, and predictions are matched to targets by Hungarian assignment.

Stage 1 uses query-wise decoder training. Image encoding and Hungarian matching are performed once per batch. The mask decoder then processes one

query per forward-backward chunk, and gradients accumulate before a single optimizer update. We also adapt the denoising principle of DN-DETR [8] to prompt-based segmentation. Each target generates perturbed mask and box prompts that pass through the prompt encoder and mask decoder. The resulting mask and class predictions are supervised against the clean target. One denoising copy is used during the first 40 epochs and two copies thereafter. Fourier augmentation perturbs frequency-domain appearance while preserving the spatial label map.

Let $m_q(v)$ denote the mask logit of query q at voxel v, and let a_{qc} denote its logit for class c. The class-specific soft prior is obtained by

$$p_c(v) = \sum_{q=1}^{Q} \frac{\exp(a_{qc})}{\sum_{c'} \exp(a_{qc'})} \, \sigma(m_q(v)) \, . \tag{1}$$

This aggregation retains both the spatial confidence of each query mask and its class confidence. The myocardium, edema, and scar maps then provide the soft priors for Stage 2.

Leakage-Free OOF Training-Input Construction. OOF prior construction creates leakage-free Stage 2 training inputs. For case i assigned to fold k, its prior is

$$P_i = f_{\theta_{-k}}(X_i), \tag{2}$$

where $f_{\theta_{-k}}$ is fitted without any case from fold k. Thus, no label from case i contributes to the model that generates its refinement prior. The first three channels contain the LGE, T2-weighted, and bSSFP images. Seven additional channels are derived from the Stage 1 probabilities:

$$[\, p_{myo},\ p_{edema},\ p_{scar},\ p_{fg},\ p_{bdry},\ p_{unc},\ p_{sdf} \,]. \tag{3}$$

Here, p_{fg} is the maximum foreground probability. The terms p_{bdry}, p_{unc}, and p_{sdf} denote an in-plane boundary map, uncertainty map, and normalized myocardium signed-distance map, respectively. The 10-channel tensor therefore combines image appearance, class confidence, geometric support, and uncertainty.

This representation makes the first-stage prediction explicit, allowing Stage 2 to learn corrections for discontinuities, implausible fragments, and uncertain boundaries.

2.3 Stage 2: Prior-Guided Bidirectional Propagation Refiner

PBPR comprises Anchor-Guided Class-Wise Propagation and 3D Residual Refinement. The propagation path combines a hybrid slice encoder, prior-derived anchor initialization, and independent bidirectional ConvGRUs. As shown in Fig. 2, it propagates myocardium, edema, and scar features independently along the slice axis. The residual refiner then adjusts the propagated foreground logits. Because all prompts are generated from Stage 1 predictions, inference remains fully automatic.

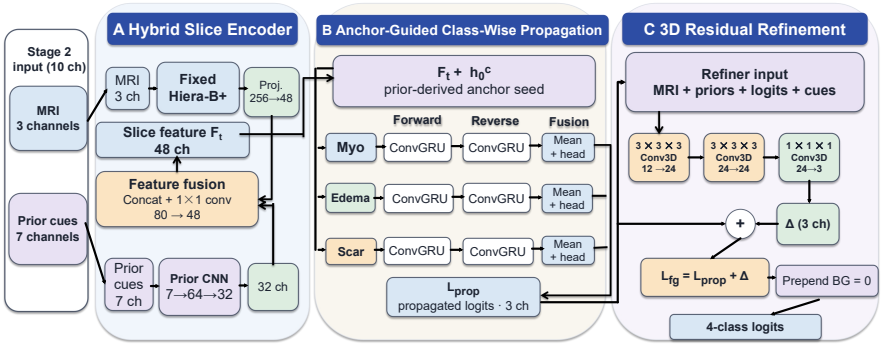

Fig. 2. The architecture of PBPR. Anchor-Guided Class-Wise Propagation combines a fixed Hiera-B+ image branch, used without a pretrained checkpoint, with a trainable prior CNN. Prior-derived anchors seed three class-wise bidirectional ConvGRUs. A full-resolution 3D residual refiner adjusts the propagated logits.

Prior-Derived Anchor Initialization. The myocardium, edema, and scar priors are analyzed separately. Candidate slices are ranked using probability mass, adaptive thresholds, and a minimum inter-anchor gap. The highest-ranked mask prompt initializes the recurrent state for each class. Additional ranked prompts contribute to anchor-reconstruction supervision, and a myocardium box prompt provides a coarse geometric constraint during training.

Hybrid Slice Encoder. The hybrid slice encoder maps the 10-channel tensor to per-slice features. Its image branch uses the SAM 2.1 Hiera-B+ architecture [12] solely as a fixed image-feature transform. No SAM 2.1 or MedSAM2 checkpoint is loaded. The randomly initialized Hiera backbone remains frozen, whereas its output projection and the prior CNN are trainable. A lightweight convolutional branch processes the seven prior-derived channels. The two streams are concatenated and projected to a shared 48-channel feature space. The resulting tensor has shape $B \times D \times 48 \times H \times W$.

Class-Wise Bidirectional ConvGRUs. Myocardium, edema, and scar use separate recurrent states, depthwise-separable ConvGRU cells, and 1×1 prediction heads. For class c, the two trajectories follow

$$h_t^{c,\rightarrow} = \mathrm{GRU}_c(F_t, h_{t-1}^{c,\rightarrow}), \tag{4}$$

$$h_t^{c,\leftarrow} = \mathrm{GRU}_c(F_t, h_{t+1}^{c,\leftarrow}), \tag{5}$$

$$\ell_{c,t}^{prop} = W_c\left(\tfrac{1}{2}\left(h_t^{c,\rightarrow} + h_t^{c,\leftarrow}\right)\right). \tag{6}$$

The forward terminal state initializes the reverse traversal. Class-specific memory allows the three structures to learn distinct through-plane transitions while retaining information from adjacent slices.

3D Residual Refinement. The propagated logits already encode cross-slice continuity, but they may still oversmooth fine boundaries or spread local errors along the depth axis. A compact 3D residual refiner is therefore appended to locally correct propagation errors and refine class boundaries. The refiner input contains 12 channels:

$$[\, I_{LGE}, I_{T2}, I_{bSSFP}, p_{myo}, p_{edema}, p_{scar}, \ell^{prop}_{myo}, \ell^{prop}_{edema}, \ell^{prop}_{scar}, p_{unc}, p_{bdry}, p_{sdf} \,]. \tag{7}$$

A full-resolution 3D residual module uses two $3 \times 3 \times 3$ convolutional layers with 24 channels and a $1 \times 1 \times 1$ decoder to predict a three-channel residual Δ. Let $L_{\text{prop}} = [\ell^{prop}_{myo}, \ell^{prop}_{edema}, \ell^{prop}_{scar}]$ denote the propagated foreground logits. Following the notation in Fig. 2, the refined foreground and final four-class logits are

$$L_{\text{fg}} = L_{\text{prop}} + \Delta, \qquad L_{\text{final}} = \text{concat}(0, L_{\text{fg}}). \tag{8}$$

The zero background logit is therefore prepended to the three refined foreground logits.

2.4 Training Objectives

Stage 1 and Stage 2 use different objectives. The Stage 1 query path follows MaskSAM's Hungarian-matched classification, mask, Dice, auxiliary-mask, and auxiliary-box terms. The box terms comprise L_1 and generalized IoU losses. Denoising forward passes add class cross-entropy, point-sampled mask, and Dice losses against clean targets.

For Stage 2, the final four-class volume is supervised by equally weighted soft-Dice and class-weighted cross-entropy terms. Three auxiliary terms encourage forward/backward consistency, anatomy-supported pathology predictions, and anchor reconstruction:

$$\mathcal{L}_2 = 0.5\mathcal{L}_{Dice} + 0.5\mathcal{L}_{CE} + \lambda_{cons}\mathcal{L}_{cons} + \lambda_{anat}\mathcal{L}_{anat} + \lambda_{anchor}\mathcal{L}_{anchor}. \tag{9}$$

We set $\lambda_{cons} = 0.005$, $\lambda_{anat} = 0.03$, and $\lambda_{anchor} = 0.02$. The auxiliary terms regularize the propagation path without adding prediction heads.

3 Experiments

3.1 Dataset and Implementation

We use the complete three-sequence MyoPS cohort from centers B and C, comprising 35 and 45 annotated cases, respectively, and report results on the official 15-case validation set. The data originate from the CARE 2026 Myocardium track [15,11,4,3]. LGE, T2-weighted, and bSSFP images are converted to nnU-Net format in that order [6]. Labels are mapped to background, myocardium, edema, and scar, and restored to the official identifiers for submission. Five-fold partitioning of the 80 training cases provides leakage-free OOF priors.

Stage 1 uses a MedSAM-initialized ViT-B encoder, a $4 \times 256 \times 256$ patch, and eight queries. Training runs for 500 epochs, with 250 training and 50 validation iterations per epoch. We use SGD with Nesterov momentum 0.99, weight decay 3×10^{-5}, and an initial learning rate of 10^{-3}. A polynomial schedule decays the learning rate. The query chunk size is one. After a 40-epoch warm-up, each target uses two noisy prompt copies. Fourier augmentation uses random frequency filtering (RFF) with $p_{\mathrm{RFF}} = 0.5$ and cross-sample amplitude mixing (CSAM) with $p_{\mathrm{CSAM}} = 0.3$.

Stage 2 is trained on 10-channel OOF tensors for 80 epochs, with 300 training and 60 validation iterations per epoch. Foreground oversampling is 0.80, and the initial learning rate is 8×10^{-4}. The optimizer, weight decay, and polynomial schedule match Stage 1. Cross-entropy class weights are $[0.5, 1.0, 2.2, 3.0]$ for background, myocardium, edema, and scar. For the propagation-only ablation, the 3D residual contribution is disabled by setting its scale to zero; the full model enables residual refinement. All other training and evaluation settings are unchanged.

4 Results

4.1 Ablation Study Results

Table 1 evaluates the incremental contributions of the Enhanced MaskSAM Prior Generator, the integrated Anchor-Guided Class-Wise Propagation path, and 3D Residual Refinement on the validation set.

Table 1. Ablation results on the official validation set. OOF input construction is fixed for all Stage 2 variants. The Anchor-Guided Class-Wise Propagation row jointly adds hybrid slice encoding, prior-derived anchor initialization, class-wise bidirectional ConvGRUs, and the fixed propagation objectives. The full model additionally enables 3D Residual Refinement. HD is reported in millimeters.

Setting	Scar Dice↑	Scar HD↓	Edema Dice↑	Edema HD↓
MaskSAM baseline	0.6809	18.6108	0.7058	23.6721
Enhanced MaskSAM Prior Generator	0.6911	13.5719	0.7168	20.2337
+ Anchor-Guided Class-Wise Propagation	0.7140	13.9073	**0.7324**	20.4347
+ 3D Residual Refinement (full model)	**0.7200**	**12.5062**	0.7322	**19.7994**

The prior generator increased scar and edema Dice by 0.0102 and 0.0110, respectively. The complete propagation-path row increased the corresponding Dice scores by 0.0229 and 0.0156; its HD changes were +0.3354 and +0.2010 mm. Residual refinement added 0.0060 scar Dice and reduced scar and edema HD by 1.4011 and 0.6353 mm, respectively; edema Dice changed by −0.0002.

Figure 3 visually compares the predicted edema and scar regions. In the displayed slices, the MaskSAM-PBPR contours appear more coherent and less

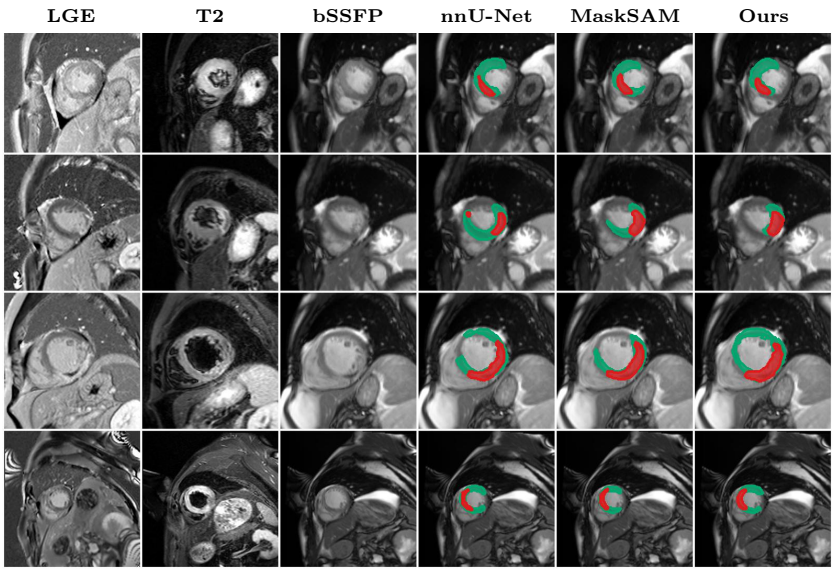

Fig. 3. Qualitative comparison on validation slices. Columns 1–3 show LGE, T2-weighted, and bSSFP inputs; columns 4–6 overlay nnU-Net, MaskSAM, and MaskSAM-PBPR predictions on bSSFP. Green and red denote edema and scar; ground truth is not shown.

fragmented than the baseline outputs. Because ground truth is not shown, this comparison describes prediction morphology only; quantitative accuracy is assessed by the Dice and HD values in Table 2.

4.2 Comparison with Baselines

We compare MaskSAM-PBPR with nnU-Net and the single-stage MaskSAM baseline on the same official validation set.

Table 2. Results on the official 15-case validation set. HD is reported in millimeters.

Method	Scar Dice↑	Scar HD↓	Edema Dice↑	Edema HD↓
nnU-Net	0.6215	26.9650	0.6706	29.5921
MaskSAM baseline	0.6809	18.6108	0.7058	23.6721
MaskSAM-PBPR	**0.7200**	**12.5062**	**0.7322**	**19.7994**

MaskSAM-PBPR outperformed both baselines on Dice and HD for both pathologies. Relative to nnU-Net, scar and edema Dice increased by 0.0985 and 0.0616, while the corresponding HD values decreased by 14.4588 and 9.7927 mm.

Relative to MaskSAM, the corresponding Dice gains were 0.0391 and 0.0264, and the HD reductions were 6.1046 and 3.8727 mm.

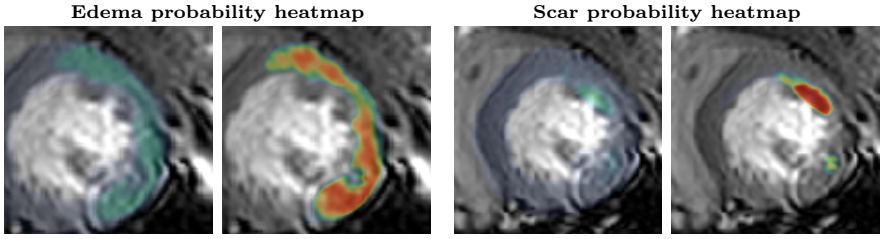

Fig. 4. Class-specific probability maps for one validation case. Each pair compares MaskSAM (left) with MaskSAM-PBPR (right); MaskSAM-PBPR shows more concentrated responses, and warmer colors indicate higher probabilities.

5 Discussion and Conclusion

The ablation supports complementary roles for prior generation, the complete Anchor-Guided Class-Wise Propagation path, and 3D Residual Refinement. The complete propagation path improved Dice for both pathologies, whereas residual refinement improved scar Dice and both HD values.

Performance under missing or misaligned sequences and on larger multi-center cohorts remains untested. Overall, MaskSAM-PBPR provides a practical two-stage route for converting query-derived soft priors into more coherent volumetric predictions.

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

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Computer Vision – ECCV 2020. Lecture Notes in Computer Science, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
2. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299 (2022)
3. Ding, W., Li, L., Qiu, J., Lin, B., Yang, M., Huang, L., Wu, L., Wang, S., Zhuang, X.: CineMyoPS: Segmenting myocardial pathologies from cine cardiac MR. IEEE Transactions on Medical Imaging **44**(12), 5027–5037 (2025). https://doi.org/10.1109/TMI.2025.3586254

4. Ding, W., Li, L., Qiu, J., Wang, S., Huang, L., Chen, Y., Yang, S., Zhuang, X.: Aligning multi-sequence CMR towards fully automated myocardial pathology segmentation. IEEE Transactions on Medical Imaging **42**(12), 3474–3486 (2023). https://doi.org/10.1109/TMI.2023.3288046
5. Gao, J., Cai, Y., Zhao, Z., Lan, X., Huang, Q., Lan, L., Li, T.Q.: RGU-Mamba: An U-Mamba network with region-based training optimized for domain generalization applied to myocardial scar and edema segmentation. In: Comprehensive Analysis and Computing of Real-World Medical Images. Lecture Notes in Computer Science, vol. 15548, pp. 77–86. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-87009-5_8
6. 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**, 203–211 (2021)
7. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
8. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: DN-DETR: Accelerate DETR training by introducing query denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13619–13627 (2022)
9. Li, L., Wu, F., Wang, S., Luo, X., Martín-Isla, C., Zhai, S., Zhang, J., Liu, Y., Zhang, Z., et al.: MyoPS: A benchmark of myocardial pathology segmentation combining three-sequence cardiac magnetic resonance images. Medical Image Analysis **87**, 102808 (2023)
10. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024). https://doi.org/10.1038/s41467-024-44824-z
11. Qiu, J., Li, L., Wang, S., Zhang, K., Chen, Y., Yang, S., Zhuang, X.: MyoPS-Net: Myocardial pathology segmentation with flexible combination of multi-sequence CMR images. Medical Image Analysis **84**, 102694 (2023)
12. Ravi, N., Gabeur, V., Hu, Y.T., et al.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
13. Xie, B., Tang, H., Duan, B., Cai, D., Yan, Y., Agam, G.: MaskSAM: Auto-prompt SAM with mask classification for volumetric medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24423–24433 (2025). https://doi.org/10.1109/ICCV51701.2025.02264
14. Zhang, X., Peng, Y., Zhang, Z.: EHU-Mamba2: Enhanced U-Mamba for multi-center cardiac MR segmentation with dynamic alignment and adaptive upsampling. In: Comprehensive Analysis and Computing of Real-World Medical Images. Lecture Notes in Computer Science, vol. 16257, pp. 225–234. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-16271-7_21
15. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(12), 2933–2946 (2019)