Structure-Reliability-Guided ROI and Donor Fusion for Multi-Center MR Whole-Heart Segmentation

Yanhong Du[2], Xiao Zhang[2⋆], Xiaoan Chen[3], Jun Feng[2], and Jingkun Chen[1⋆]

[1] School of Automation, Northwestern Polytechnical University, Xi'an, China
jingkunchen@nwpu.edu.cn
[2] College of Computer Science, Northwest University, Xi'an, China
xiaozhang@nwu.edu.cn
[3] Affiliated Hospital of Northwest University (Xi'an No. 3 Hospital), Xi'an, China

Abstract. Multi-center MR whole-heart segmentation is challenging because cardiac chambers, myocardium, and great vessels have different sizes, boundary contrasts, and field-of-view coverage. nnU-Net provides a competitive baseline, but our validation analysis shows that the remaining errors concentrate in myocardium (Myo), aorta (AO), and pulmonary artery (PA). We address these structure-specific challenges with a structure-reliability-guided inference framework. A five-fold nnU-Net prediction is used as the anatomical anchor, AO and Myo are refined by an automatic second-stage ROI branch, and an MAE-pretrained ResEncL nnU-Net with lightweight GraTa-style test-time adaptation serves as a donor for LV, RV, Myo, and PA. The donor assignment is determined from out-of-fold (OOF) validation and fixed before official inference. Donor predictions are accepted only through an add-only rule, while AO, LA, and RA are locked after ROI tuning. On internal five-fold OOF validation, the final pipeline improved mean Dice from 0.8843 to 0.8864. The largest gains were observed on PA (+0.0053), Myo (+0.0038), AO (+0.0027), and RV (+0.0017). On the CARE 2026 official MR validation leaderboard snapshot, the method achieved 0.8668 Dice, 14.3296 HD, and 1.6440 ASSD, ranking first at the recorded submission time. These results suggest that MR whole-heart segmentation can benefit from assigning model responsibility at the anatomical-structure level.

Keywords: MR Whole-Heart Segmentation · Region-of-Interest Refinement · Masked Autoencoder Pre-Training · Test-Time Adaptation · Selective Fusion

1 Introduction

Whole-heart segmentation aims to delineate the left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), myocardium (Myo), aorta (AO),

⋆ Corresponding author.

and pulmonary artery (PA). It is a fundamental task in cardiac image analysis and has been extensively studied in whole-heart segmentation benchmarks [19]. Early encoder-decoder and volumetric CNN architectures established the foundation of modern biomedical image segmentation [4, 9, 10]. Recent weak-supervision studies have also addressed annotation cost, label inconsistency, and human-attention cues in medical image segmentation [1, 2, 17]. Compared with single-organ tasks, whole-heart segmentation requires simultaneous accuracy on large chambers, thin myocardium, and vessels with variable visible extent.

MR whole-heart segmentation is particularly challenging under multi-center variations in intensity, contrast, and field-of-view coverage. Although modern self-configuring and architecture-scaled segmentation frameworks have substantially improved overall performance [6, 8, 11], their residual errors are often unevenly distributed across anatomical structures. In our MR validation analysis, chamber structures were relatively stable, whereas Myo, AO, and PA remained more difficult. This structure-wise performance heterogeneity becomes important when overall segmentation performance is already high, and is related to the broader challenge of learning under imbalanced anatomical targets [12]: a model or refinement strategy that improves one structure may provide little benefit, or even degrade performance, for another. Consequently, globally selecting or uniformly averaging models may be suboptimal when the remaining errors are structure dependent.

This observation motivates a structure-level reliability perspective, in which each auxiliary model is assigned only to the anatomical structures where it provides validated complementary benefits. Cascaded ROI refinement is useful for local difficult structures, while self-supervised pre-training and test-time adaptation can offer complementary predictions under domain shift [3, 7, 13, 15, 16]. However, these auxiliary predictions need not be equally reliable across all anatomical structures. This motivates using structure-level reliability to determine where each auxiliary branch should contribute.

We propose an MR-specific structure-reliability-guided fusion pipeline, in which class-wise reliability is estimated from OOF validation. The first component is a second-stage AO/Myo ROI refinement branch, which focuses on local structures that are difficult for the global model. The second component is an MAE-pretrained donor with lightweight GraTa-style adaptation, which is used only for structures where out-of-fold (OOF) validation indicates complementary benefit. The final prediction is produced by an add-only fusion rule and class-aware post-processing.

The contributions of this paper are threefold:

1. We formulate MR whole-heart refinement as a structure-level reliability assignment problem.
2. We combine AO/Myo second-stage ROI refinement with a restricted MAE+GraTa-lite donor, allowing only LV, RV, Myo, and PA supplementation while locking AO, LA, and RA.
3. We validate the design using five-fold OOF ablation and official validation results, showing overall improvements over an nnU-Net MR baseline.

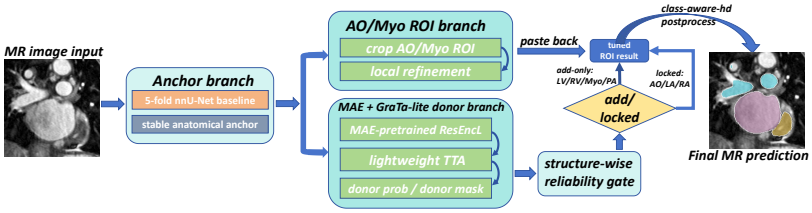

Fig. 1. Overview of the proposed MR pipeline. A five-fold nnU-Net prediction is used as the anatomical anchor. AO and Myo are refined by an automatic ROI branch. An MAE-pretrained donor with GraTa-lite produces complementary candidates, but fusion is restricted to LV, RV, Myo, and PA. AO, LA, and RA are locked after ROI tuning.

2 Method

2.1 Overview

Fig. 1 summarizes the proposed MR inference pipeline. The five-fold nnU-Net prediction is used as the anatomical anchor, and auxiliary models are introduced as structure-specific donors. Therefore, the final prediction is obtained through structure-wise refinement and selective fusion rather than global model replacement or uniform model averaging.

Given an MR image I, the pipeline first obtains an anchor prediction by averaging the softmax outputs of five nnU-Net folds. The anchor prediction is refined for AO and Myo using an automatic ROI branch. In parallel, an MAE-pretrained ResEncL nnU-Net with lightweight GraTa-style adaptation generates candidate predictions for selected donor structures. A fixed reliability assignment derived from OOF validation determines which structures are allowed to receive donor supplementation. The final output is obtained by add-only fusion, followed by MR-specific class-aware post-processing and official label restoration.

2.2 Five-Fold Anchor and Structure Reliability Assignment

The baseline anchor is produced by five-fold softmax averaging over the label set $\mathcal{L} = \{LV, RV, LA, RA, Myo, AO, PA\}$:

$$P^{anc}(x,c) = \frac{1}{K}\sum_{k=1}^{K} P_k(x,c), \quad K = 5, \tag{1}$$

where $P_k(x,c)$ is the probability predicted by fold k for class c at voxel x, and Y^{anc} is obtained by argmax. We define structure reliability as an OOF-estimated class-wise trust assignment rather than a case-wise oracle. For donor model d and class c, the OOF gain is

$$\Delta_{d,c}^{OOF} = \mathrm{DSC}_{d,c}^{OOF} - \mathrm{DSC}_{main,c}^{OOF}, \tag{2}$$

A donor is allowed for a structure only when OOF validation indicates a positive gain and no adjacent stable structure is degraded. This assignment is fixed before official inference and does not use validation or test labels at inference time.

Based on the OOF analysis, AO and Myo are handled by the ROI branch, while the MAE+GraTa-lite donor is allowed only for LV, RV, Myo, and PA. AO, LA, and RA are locked after ROI tuning:

$$\mathcal{C}_{donor}^{MR} = \{LV, RV, Myo, PA\}, \quad \mathcal{C}_{lock}^{MR} = \{AO, LA, RA\}. \tag{3}$$

This converts model fusion from global model selection into class-level responsibility assignment.

2.3 Automatic AO/Myo ROI Refinement

The ROI branch targets AO and Myo because these structures were among the lowest-performing classes in the anchor prediction. The ROI is generated automatically from the first-stage anchor prediction. For each target class $c \in \{AO, Myo\}$, we compute a 3D bounding box around the post-processed anchor component and enlarge it by 20 mm, with a minimum margin of 8 voxels along each axis. The cropped patch is forwarded to a second-stage local model trained with binary ROI supervision. Separate local models are trained for AO and Myo.

During inference, the crop is generated only from the first-stage prediction, without using ground-truth labels. OOF evaluation uses the held-out ROI fold, while official inference uses five ROI folds with mirroring test-time augmentation. The local prediction is restored to the original image space, and only AO and Myo voxels can be updated:

$$Y_c^{roi} = R_c(Y^{anc}, I), \quad c \in \{AO, Myo\}, \tag{4}$$

where R_c denotes crop, local refinement, and paste-back for class c. In the tuned setting, both AO and Myo use a probability threshold of 0.85. Added AO voxels must lie within 4 mm of the anchor AO region, and added Myo voxels must lie within 3 mm of the anchor Myo region. Other anatomical labels are protected from being overwritten, so the second stage acts as a local corrector rather than a whole-heart replacement model.

2.4 MAE-Pretrained Donor with GraTa-Lite Adaptation

The second auxiliary source is an MAE-pretrained ResEncL nnU-Net donor. MAE pre-training learns volumetric structural representations through masked reconstruction [7, 15], which can provide complementary predictions under center and intensity shift. However, OOF analysis showed that this donor should not globally replace the tuned ROI result. We therefore use it only as a restricted donor for validated structures.

To improve robustness at inference time, we apply a lightweight GraTa-style adaptation to the donor. For each case and fold model, checkpoint weights are restored before adaptation, so updates do not accumulate. We update only normalization affine parameters for one AdamW step with learning rate 10^{-5}, zero weight decay, and maximum gradient norm 1.0. The self-supervised objective is

$$\mathcal{L}_{tta} = \mathcal{L}_{ent}(P_\theta(I)) + \lambda \mathcal{L}_{con}(P_\theta(a_w(I)), P_\theta(a_s(I))), \tag{5}$$

where $\lambda = 1.0$, $\mathcal{L}_{ent}$ is entropy loss, and $\mathcal{L}_{con}$ is the mean squared consistency loss between the strong prediction and the stop-gradient weak prediction. The weak augmentation a_w is the original preprocessed patch. The strong augmentation a_s is intensity-only, with brightness multiplier $1 + \mathcal{N}(0, 0.02)$ and Gaussian noise of 0.01 times the patch intensity standard deviation. No spatial transformation is used. The update is applied only when the entropy and consistency gradients have non-negative alignment. The adapted parameters are discarded after each case, and the branch only supplies candidate voxels for $\mathcal{C}^{MR}_{donor}$.

2.5 Locked-Class Add-Only Fusion and MR Post-Processing

Let Y^{main} denote the tuned AO/Myo ROI result and Y^{donor} denote the MAE+GraTa-lite donor prediction. For a donor-allowed class $c \in \mathcal{C}^{MR}_{donor}$, donor voxels are accepted only where the main prediction is not locked:

$$Y^{final}_c = Y^{main}_c \cup \{x : Y^{donor}(x) = c,\, Y^{main}(x) \notin \mathcal{C}^{MR}_{lock}\}. \tag{6}$$

For locked classes, the tuned ROI result is preserved:

$$Y^{final}_c = Y^{main}_c, \quad c \in \mathcal{C}^{MR}_{lock}. \tag{7}$$

This rule prevents donor predictions from overwriting AO, LA, and RA. Myo appears in both auxiliary branches, but the donor can only supplement Myo; it cannot remove the tuned ROI result. After fusion, MR-specific class-awarehd post-processing suppresses isolated components that may increase surface-distance errors. The final prediction is restored to the official label values and saved in the original image space.

3 Experiment

3.1 Dataset, Setup, and Metrics

Experiments were conducted on the MR part of the CARE 2026 whole-heart track, which builds upon prior whole-heart and myocardial segmentation resources and methodologies [5, 18, 20]. The training set contains 46 MR volumes from centers C, D, and E, and the official validation set contains 20 MR volumes. The seven foreground labels are LV, RV, LA, RA, Myo, AO, and PA. Internal evaluation uses five-fold OOF predictions, while official validation scores are taken from the leaderboard snapshot.

We report Dice similarity coefficient (DSC), Hausdorff distance (HD), and average symmetric surface distance (ASSD). DSC measures volume overlap, while HD and ASSD measure boundary errors in physical space [14]. The mean DSC is the average over the seven foreground structures:

$$\overline{\mathrm{DSC}} = \frac{1}{7} \sum_{c \in \mathcal{L}} \mathrm{DSC}_c. \tag{8}$$

Table 1. MR five-fold OOF ablation of the proposed pipeline. Class-wise DSC values are rounded to four decimals. Bold indicates the best or tied-best rounded value among the proposed-pipeline variants; lower is better for HD and ASSD.

	Class-wise DSC							Mean	Surface distance	
Method	LV	RV	LA	RA	Myo	AO	PA	DSC	HD	ASSD
nnU-Net 5-fold + class-aware-hd	0.9455	0.9061	**0.9101**	**0.9233**	0.8586	0.8235	0.8234	0.8843	16.1886	1.9518
Baseline + AO/Myo ROI refinement	0.9455	0.9061	**0.9101**	**0.9233**	0.8610	**0.8262**	0.8234	0.8850	16.1616	1.9610
MAE-pretrained donor + class-aware-hd	**0.9472**	0.9073	0.9096	0.9231	0.8605	0.8258	0.8253	0.8855	**15.5870**	1.9514
Tuned ROI + MAE add-only fusion	0.9466	**0.9078**	**0.9101**	**0.9233**	**0.8625**	**0.8262**	0.8286	**0.8864**	15.9000	**1.9318**
Final: tuned ROI + MAE/GraTa-lite add-only fusion	0.9466	**0.9078**	**0.9101**	**0.9233**	0.8624	**0.8262**	**0.8287**	**0.8864**	15.8968	**1.9318**

3.2 OOF Ablation

Table 1 reports the MR OOF ablation for the proposed structure-reliability-guided pipeline. The baseline obtained 0.8843 mean DSC. The lower classes were Myo (0.8586), AO (0.8235), and PA (0.8234), confirming that the residual errors were concentrated in thin myocardium and vessels. AO/Myo ROI refinement increased Myo to 0.8610 and AO to 0.8262. The MAE donor improved several classes, but its class-wise behavior differed from the ROI-tuned anchor. The final add-only design achieved 0.8864 mean DSC. The largest class-wise gains over the baseline were observed on PA, Myo, AO, and RV, while LA and RA remained unchanged because they were locked during donor fusion.

To directly examine global model averaging, we additionally evaluated a post-hoc 0.5/0.5 probability average between the anchor and MAE donor (Table 2). This check is reported separately from the proposed-pipeline ablation because it is a global-fusion baseline rather than a component of our submitted method. Uniform averaging produced a comparable mean DSC, but it reduced the two targeted difficult structures, Myo and PA, compared with the final selective pipeline. We therefore use this result to support a conservative conclusion: global averaging can shift structure-wise performance, whereas the selective pipeline keeps responsibility fixed to the validated structures.

Table 2. Separate post-hoc comparison between the submitted selective pipeline and global probability averaging. Uniform averaging is not part of the proposed pipeline.

Method	Mean DSC	Myo DSC	PA DSC
Final selective pipeline	0.8864	**0.8624**	**0.8287**
Uniform avg. anchor+MAE donor	0.8869	0.8617	0.8271

3.3 Official Validation Results

Table 3 summarizes our official MR validation submissions. The stable baseline achieved 0.8641 DSC. Tuned AO/Myo ROI refinement increased DSC to 0.8651. The final pipeline further increased DSC to 0.8668 and reduced both HD and ASSD. This trend is consistent with the internal OOF ablation and supports the transferability of the structure-level assignment.

3.4 Qualitative Comparison

Fig. 2 shows representative OOF predictions across refinement stages. With the same anatomical color map, the comparison illustrates that ROI refinement mainly adjusts local AO/Myo regions, while the final donor-fusion stage complements selected structures without visibly disturbing stable chambers.

Table 3. Official MR validation leaderboard snapshot for our submissions. Bold indicates the best value among our listed submissions; lower is better for rank, HD, and ASSD.

Submission	Rank	CreatedTime	DSC	HD	ASSD
Stable MR baseline	6	2026-06-08 14:09:07	0.8641	14.4985	1.6557
Tuned AO/Myo ROI	3	2026-06-25 17:29:50	0.8651	14.5039	1.6614
Final MR pipeline	**1**	2026-07-02 13:52:37	**0.8668**	**14.3296**	**1.6440**

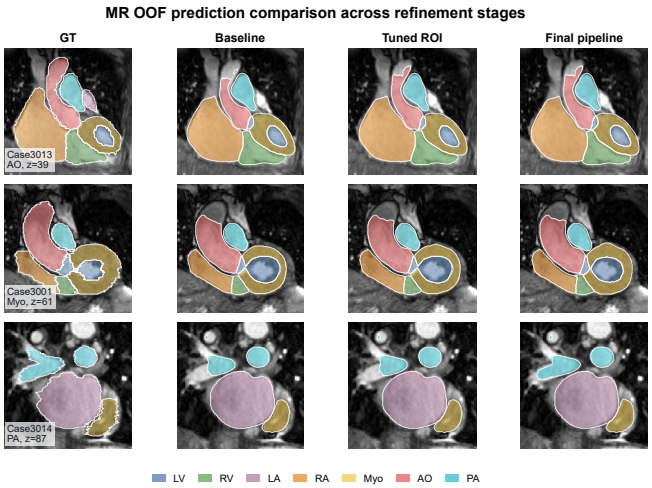

Fig. 2. Representative MR OOF prediction comparison across refinement stages. Columns show ground truth, nnU-Net baseline, tuned AO/Myo ROI, and final prediction.

4 Conclusion

We proposed an MR-specific structure-reliability-guided inference framework for whole-heart segmentation. The method uses AO/Myo ROI refinement to adjust local difficult structures and introduces an MAE+GraTa-lite donor only for validated complementary classes. Add-only fusion and locked-class rules prevent donor predictions from overwriting protected anatomical structures. Internal OOF and official validation results show overall improvements over an nnU-Net baseline, suggesting that structure-level model responsibility is a useful strategy for MR whole-heart segmentation.

Future work will explore how to estimate structure-level reliability more adaptively across datasets and institutions. In the current study, the reliability assignment is fixed before inference based on OOF validation; incorporating case- or domain-specific reliability cues may further improve transferability to new acquisition settings.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No.62403380).

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

References

1. Chen, J., Duan, H., Zhang, X., Gao, B., Grau, V., Han, J.: From gaze to insight: Bridging human visual attention and vision-language model explanation for weakly-supervised medical image segmentation. IEEE Transactions on Medical Imaging (2025)
2. Chen, J., Huang, W., Zhang, J., Debattista, K., Han, J.: Addressing inconsistent labeling with cross image matching for scribble-based medical image segmentation. IEEE Transactions on Image Processing **34**, 842–853 (2025)
3. Chen, Z., Ye, Y., Pan, Y., Xia, Y.: Gradient alignment improves test-time adaptation for medical image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2429–2437 (2025)
4. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: Learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention. pp. 424–432. Springer (2016)
5. Gao, S., Zhou, H., Gao, Y., Zhuang, X.: BayeSeg: Bayesian modeling for medical image segmentation with interpretable generalizability. Medical Image Analysis **89**, 102889 (2023)
6. Hatamizadeh, A., Yang, D., Roth, H., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)
7. He, K., Chen, X., Xie, S., Li, Y., Dollar, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
8. 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)
9. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision. pp. 565–571 (2016)
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
11. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: Mednext: Transformer-driven scaling of convnets for medical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. pp. 405–415. Springer (2023)
12. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 240–248. Springer (2017)
13. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A.A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: International Conference on Machine Learning. pp. 9229–9248 (2020)
14. Taha, A.A., Hanbury, A.: Metrics for evaluating 3d medical image segmentation: Analysis, selection, and tool. BMC Medical Imaging **15**, 29 (2015)
15. Wald, T., Ulrich, C., Lukyanenko, S., Goncharov, A., Paderno, A., Maerkisch, L., Jaeger, P.F., Maier-Hein, K.H.: Revisiting mae pre-training for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)

16. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (2021)
17. Zhang, X., Wu, S., Zhang, P., Jin, Z., Xiong, X., Bu, Q., Chen, J., Feng, J.: HELPNet: Hierarchical perturbations consistency and entropy-guided ensemble for scribble supervised medical image segmentation. Medical Image Analysis **105**, 103719 (2025). https://doi.org/10.1016/j.media.2025.103719
18. 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)
19. Zhuang, X., Li, L., Payer, C., Stern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, Ö., Bian, C., et al.: Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge. Medical Image Analysis **58**, 101537 (2019)
20. Zhuang, X., Shen, J.: Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. Medical Image Analysis **31**, 77–87 (2016)