Hybrid-Supervised Multi-Task Network for Multi-Modal Perioperative Mitral Segmentation

Yue Qi[1], Zheng Li[2], Junxuan Yu[3], and Xin Yang[3⋆]

[1] College of Chemistry and Life Science, Beijing University of Technology, Beijing, China
[2] Independent Researcher, Tianjin, China
[3] School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China

Abstract. Multi-modal perioperative mitral valve segmentation for cardiac CT, transesophageal echocardiography (TEE), and surgical endoscopy is critical for cardiac interventional planning, yet faces tissue blurring, heavy instrument occlusion, and large cross-modality domain gaps. This work proposes a hybrid-supervised multi-task framework with modality-adaptive subnetworks for these heterogeneous modalities, integrating semi-supervised consistency learning, foundation-model adaptation, and boundary-aware refinement. On the MVAA 2026 hidden test set, our method achieves DSC/HD/ASD of 0.827/5.245 mm/0.360 for CT, 0.814/11.761 mm/0.769 for TEE, and 0.847/181.592 pixel/60.443 for surgical frames. We observe substantial improvements over the official MVAA-Baseline on CT and surgical frames, alongside moderate performance gains on TEE, offering a viable alternative to heavily tuned nnU-Net v2 solutions. Qualitative results show anatomically plausible and smooth mitral-valve contours, though surgical-frame segmentation exhibits elevated HD values under heavy instrument occlusion. Our code is available at `https://github.com/Daheilou/MICCAI2026_MVAA`.

Keywords: multi-modal imaging, semi-supervised learning, mitral segmentation, Vista3D SegNet, DINOv2-DPT

1 Introduction

Precise mitral valve morphological analysis is essential for diagnosis, risk stratification and minimally-invasive intervention of structural heart disease[1]. Perioperative cardiac assessment involves three heterogeneous imaging modalities: cardiac CT, transesophageal echocardiography (TEE), and surgical endoscopy[2]. Preoperative CT provides high-resolution volumetric anatomy for surgical planning, and accurate mitral leaflet-annulus segmentation is critical for TEER and annuloplasty[3]. Intraoperative TEE captures real-time valve morphology yet suffers from acoustic artifacts and low tissue contrast[4]. Surgical frames offer surgeon-view visual evidence for intra-procedural inspection[5,6].

⋆ Corresponding author. Email: xinyang@szu.edu.cn

Nevertheless, automatic mitral segmentation for this multi-modal perioperative setting remains challenging. Manual annotation is labor-intensive and suffers from inter-observer variation[7]. As highlighted by recent MICCAI benchmarks for peri-procedural medical imaging[8,9], existing deep-learning solutions mostly target single-modality inputs and lack cross-domain generalization. They struggle to cope with distinct imaging appearances of CT, noisy TEE volumes and instrument-occluded endoscopic scenes, yielding unreliable predictions that hinder clinical usage[10].

To tackle these challenges, this work presents a hybrid-supervised multi-task framework with modality-adaptive subnetworks for peri-operative mitral segmentation across CT, TEE and surgical frames. Semi-supervised learning is utilized for annotation-scarce CT and endoscopic data, whereas TEE is trained under full supervision. Modality-specific backbones (Vista3D SegNet for 3D volumes, DINOv2-DPT for 2D surgical frames) are coupled with UniMatch-based EMA teacher-student regularization [11,12,13,14]. Experiments yield improved segmentation accuracy and anatomical consistency for automated peri-interventional mitral valve analysis.

2 Related Work

Mitral valve segmentation has been predominantly tackled using single-modality supervised pipelines [7,16]. While nnU-Net variants achieve strong results on the MVAA benchmark [21,22,23], they face three key limitations in perioperative multi-modal settings. First, annotation scarcity in CT and surgical frames limits purely supervised approaches, and existing semi-supervised frameworks generalize poorly from 2D natural images to 3D volumetric data. Second, large domain gaps between high-resolution CT, noisy TEE, and instrument-occluded surgical scenes challenge generic backbones; although foundation models like Vista3D [11] and DINOv2 [12] provide transferable priors, their adaptation remains underexplored. Third, thin leaflets demand fine-grained modeling, yet boundary-aware losses are seldom integrated into semi-supervised pipelines for this task. To address these gaps, we propose a hybrid-supervised multi-task framework, detailed in the next section.

3 Methodology

3.1 3D Mitral Valve Segmentation Pipeline for Cardiac CT

UniMatch EMA Teacher-Student Semi-Supervised Training Framework As illustrated in Figure 1, we extend the original 2D UniMatch paradigm to 3D cardiac CT for mitral valve segmentation. The teacher network is derived via EMA from the student Vista3D SegNet with no separate parameter initialization.For labeled samples, heavily augmented image-mask pairs are fed to the student for supervised loss $\mathcal{L}_{\text{sup}}$. Unlabeled volumes undergo weak-strong dual augmentation: strongly augmented volumes are forwarded to the student,

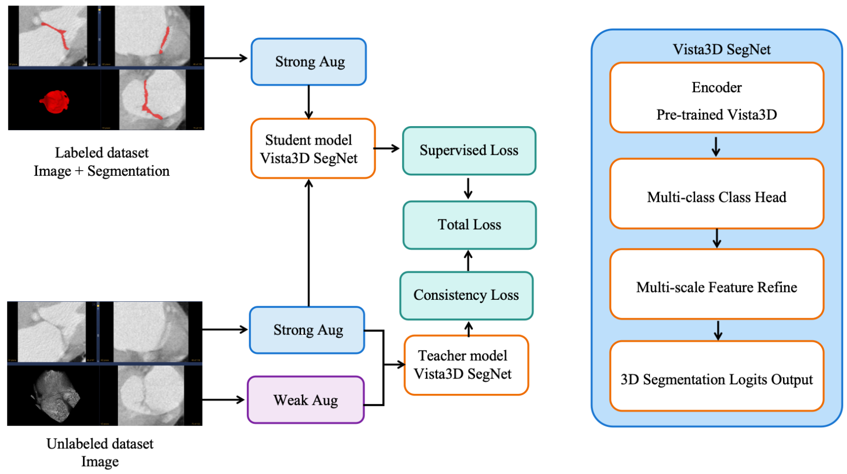

Fig. 1: Architecture of the semi-supervised 3D mitral valve segmentation pipeline for cardiac CT using UniMatch. Student and teacher branches share the Vista3D SegNet backbone with pre-trained encoder, multi-scale refinement module and segmentation head.

whereas weakly augmented ones generate soft pseudo-labels via the frozen EMA teacher. Consistency loss $\mathcal{L}_{\text{consist}}$ enforces prediction agreement. The total loss is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \lambda \cdot \mathcal{L}_{\text{consist}}, \tag{1}$$

where λ linearly ramps from 0 to 0.5. The student is updated via back-propagation, and the teacher is updated by exponential moving-average:

$$\theta_{\text{tea}} \leftarrow \alpha \cdot \theta_{\text{tea}} + (1 - \alpha) \cdot \theta_{\text{stu}}, \tag{2}$$

with $\alpha = 0.995$. EMA updates activate only during semi-supervised stages.

Originally designed for 2D tasks, UniMatch is adapted to 3D volumetric CT inputs. Weak augmentation includes single-axis random flip and low-magnitude Gaussian noise; strong augmentation comprises multi-axis flip, volumetric gamma transform, 3D Gaussian noise, volume cutout and 3D Gaussian smoothing. A supervised warm-up stage uses only labeled data while disabling unsupervised loss and EMA. After warm-up, the teacher is initialized from student weights, and EMA updates take effect.

Vista3D-base Backbone and Loss Function Design Vista3D-base serves as the volumetric backbone for our full segmentation architecture Vista3D SegNet. A multi-class segmentation head outputs logits for mitral valve structures, accompanied by a lightweight frequency-decomposition-based multi-scale feature refinement module. It refines coarse predictions by fusing low-frequency global

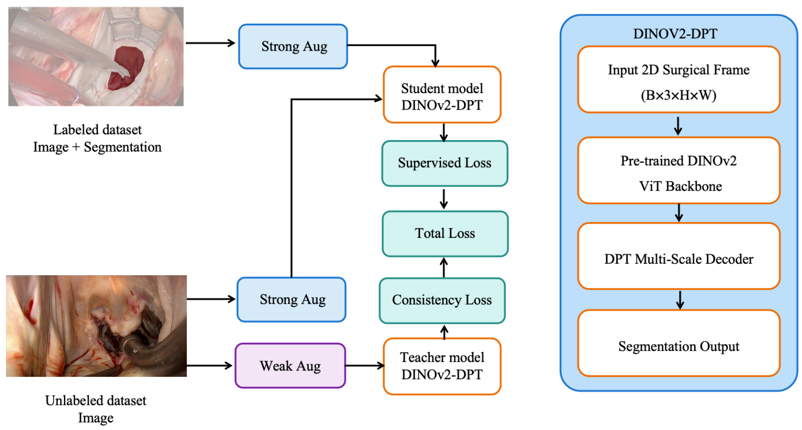

Fig. 2: Overall EMA teacher-student semi-supervised training pipeline of the proposed DINOv2-DPT mitral valve segmentation model for surgical frames.

context and high-frequency boundary features through a weighted residual:

$$h_{\text{pred}} = h_{\text{raw}} + w_{\text{res}} \cdot (\text{LowFeature} + \text{HighFeature}), \tag{3}$$

where $w_{\text{res}} = 0.2$.

For improved segmentation of thin mitral leaflets, the supervised loss combines weighted Dice-CE and gradient-aware boundary regularization:

$$\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{DiceCE}}(h_{\text{pred}}, h_{\text{gt}}) + w_{\text{bd}} \cdot \mathcal{L}_{\text{Boundary}}(h_{\text{pred}}, h_{\text{gt}}), \tag{4}$$

with $\lambda_{\text{dice}} = 0.8$, $\lambda_{\text{ce}} = 0.2$ and $w_{\text{bd}} = 0.7$.

For semi-supervised consistency regularization, low-confidence voxels from teacher pseudo-labels are masked by threshold $\tau = 0.6$. The consistency loss mixes L2 and Dice constraints over masked voxels:

$$\mathcal{L}_{\text{consist}} = \mathbb{E}\left[\frac{\sum \boldsymbol{m} \cdot \left(\lambda_{\text{ce}} \|h_{\text{stu}} - h_{\text{tea}}\|_2^2 + \lambda_{\text{dice}} \mathcal{L}_{\text{Dice}}(h_{\text{stu}}, h_{\text{tea}})\right)}{\sum \boldsymbol{m} + \epsilon}\right], \tag{5}$$

where $\lambda_{\text{dice}} = \lambda_{\text{ce}} = 0.5$ and $\epsilon = 10^{-8}$.

3.2 Fully Supervised 3D Mitral Valve Segmentation Pipeline for Intraoperative TEE

We employ the identical Vista3D SegNet architecture for intraoperative TEE mitral segmentation. Distinct from the semi-supervised UniMatch pipeline for CT, this branch follows fully supervised training without consistency regularization. This pipeline adopts Dice+CE loss.

3.3 Semi-Supervised 2D Mitral Valve Segmentation for Surgical Video Frames

As shown in Figure 2, DINOv2-DPT acts as the shared backbone for the trainable student and slowly updated EMA teacher within our semi-supervised pipeline.The input 3-channel surgical frames are fed into the pre-trained DINOv2 ViT encoder to extract multi-scale embeddings. These features are further processed by the DPT decoder to aggregate global semantics and fine anatomical boundary information for segmentation prediction. Dedicated branches output segmentation logits to support consistency regularization between student and teacher.Endoscopic data suffers from severe foreground-background imbalance, uneven illumination, motion blur and instrument occlusion. We adopt a compound loss combining Dice loss, cross-entropy loss and boundary regularization to enhance segmentation performance on thin mitral leaflets. Under the EMA teacher-student framework, we leverage strong augmentations for labeled samples, and weak–strong dual augmentation for unlabeled frames to generate pseudo-labels and compute masked consistency loss.

4 Experiments

Evaluations are conducted on the multi-modal MVAA challenge dataset [15], comprising three subtasks: CT, TEE, and surgical frames (Table 1). For CT and TEE, Vista3D SegNet initialized with Vista3D-base weights is trained for 200 epochs using AdamW ($lr = 1 \times 10^{-4}$) and cosine annealing. For surgical frames, DINOv2-DPT built upon DINOv2 ViT-base is trained with semi-supervised AdamW with linear warm-up, cosine annealing, and early-stopping. Performance is measured by DSC, HD, and ASD, where HD and ASD are reported in millimetres for volumetric tasks and in pixels for 2D frames.

Table 1: Data Partition Statistics of Three Subtasks

Subtask	Labeled Train	Unlabeled Train	Validation
Task 1 (CT)	27	1040	30
Task 2 (TEE)	105	0	20
Task 3 (Surgical Frame)	180	1379	48

5 Results

5.1 Comparison with Competing Methods

As shown in Table 2, leading submissions primarily adopt well-tuned nnU-Net v2 as the 3D backbone, yielding compelling quantitative results with limited

Table 2: Experimental results on the hidden test set of the MVAA 2026 benchmark

Method	Task 1 (CT)			Task 2 (TEE)			Task 3 (Surgical Frame)			RunTime (s)
	DSC ↑	HD(mm) ↓	ASD ↓	DSC ↑	HD(mm) ↓	ASD ↓	DSC ↑	HD(pixel) ↓	ASD ↓	
3D UNet[16]/ 3D UNet[16]/ UNet++[17]/ MVAA-Baseline[18]	0.590	12.140	1.346	0.758	17.853	1.377	0.572	555.778	302.770	115.4
3D Unet[16]// SegResNet[19]// Unet++[17] neu415 [20]	0.812	7.490	0.630	0.825	11.940	0.746	0.741	547.147	398.698	812.5
nnU-Net v2[21]/ nnU-Net v2[21]/ Unet++[17] xuhui [22]	0.841	5.327	0.478	0.852	9.043	0.563	0.801	278.822	61.115	1980.1
nnU-Net v2[21]/ nnU-Net v2[21]/ DINOv3-UNet[23] Fitz [23]	0.854	5.068	0.341	0.900	6.148	0.295	0.850	162.087	36.138	6517.5
Vista3D SegNet (Fusion)[11]/ Vista3D SegNet[11]/ DINOv2-DPT[12] **Our Method**	**0.827**	**5.245**	**0.360**	**0.814**	**11.761**	**0.769**	**0.847**	**181.592**	**60.443**	**3342.2**

Table 3: Ablation results for Task 1 (CT) validation set. Baseline: supervised-only Vista3D SegNet with Dice+CE loss (loss ratio 8:2). All models use sliding-window inference. TTA and fusion are inference-stage strategies.

Model Configuration	DSC ↑	HD(mm) ↓	ASD ↓
Vista3D SegNet	0.836	7.144	0.346
+ UniMatch SSL	0.840	5.925	0.310
+ Multi-scale feature refinement module	0.846	5.091	0.295
+ Boundary composite loss	0.845	4.871	0.296
+ TTA	0.849	4.639	0.290
+ **Fusion**	**0.851**	**4.603**	**0.289**

architectural diversity. MVAA baselines based on 3D UNet and UNet++ suffer from higher HD and ASD on noisy clinical data. On the hidden test set, our fusion-enhanced Vista3D SegNet achieves balanced CT segmentation performance. For TEE, Vista3D SegNet produces reasonable predictions yet remains inferior to heavily-optimized nnU-Net v2 variants. Our DINOv2-DPT achieves a competitive DSC of 0.847 for surgical frames, marginally below the top-ranked 0.850. Overall, the proposed pipeline offers an alternative architectural choice for mitral-valve segmentation.

5.2 Ablation

Table 3 reports ablation results on the CT validation set to validate our proposed components. Starting from the fully-supervised Vista3D SegNet baseline (DSC=0.836, HD=7.144, ASD=0.346), each added module improves segmentation accuracy and boundary quality overall. UniMatch-based semi-supervised learning exploits unlabeled CT data, lifting DSC to 0.840 and reducing HD to 5.925. The multi-scale feature refinement module further enhances fine-grained anatomical representation, bringing HD down to 5.091 and ASD to 0.295. The boundary-aware composite loss augments the original 8:2 Dice-CE objective with an auxiliary boundary-loss term scaled by 0.7. It reduces HD for sharper anatomical contours for thin mitral-leaflet structures, albeit with minor fluctuations in DSC and ASD. Test-time augmentation stabilizes inference outputs, and checkpoint fusion aggregates complementary model weights to achieve the best overall performance (DSC=0.851, HD=4.603, ASD=0.289).

Table 4: Ablation results for Task 2 (TEE) validation set. Baseline: Vista3D SegNet with Dice+CE loss (loss ratio 8:2).

Model Configuration	DSC ↑	HD(mm) ↓	ASD ↓
Vista3D SegNet	0.786	17.164	1.142
+ Multi-scale feature refinement module	0.808	15.636	0.846

Table 4 reports ablation results on the TEE validation set. Equipped with the multi-scale feature refinement module, Vista3D SegNet obtains performance gains: DSC improves from 0.786 to 0.808, HD decreases from 17.164 to 15.636, and ASD drops from 1.142 to 0.846, which verifies the effectiveness of the multi-scale refinement module for TEE segmentation.

Table 5: Ablation results for Task 3 (Surgical frames). Baseline: DINOv2-DPT trained with Dice-Focal loss (loss ratio 7:3), with TTA enabled at inference. Fusion denotes 1:1 weight-averaging of two trained models. Post-processing is an inference-stage operation.

	Validation set			Test set		
Model Configuration	DSC ↑	HD(pixel) ↓	ASD ↓	DSC ↑	HD(pixel) ↓	ASD ↓
DINOv2-DPT	0.803	101.133	12.709	–	–	–
+ SSL (w/o Conf-Mask)	0.799	14.310	14.602	–	–	–
+ **SSL (w/ Conf-Mask)**	**0.815**	**69.132**	**11.712**	**0.847**	**181.592**	**60.443**
+ Fusion	0.817	70.666	12.040	0.859	216.138	97.968
+ Post-processing	0.816	70.008	11.887	0.859	215.982	97.873

Table 5 presents ablation results of the surgical-frame subtask over validation and test splits. Weight-averaged checkpoint fusion achieves slight DSC gains but worsens HD and ASD boundary metrics, whereas post-processing (largest connected-component filtering) causes negligible performance shifts. Such a trade-off indicates that weight-based checkpoint averaging softens distinct anatomical contours, improving overlap-oriented DSC while impairing distance-sensitive metrics for thin mitral leaflets.

5.3 Visual Analysis

Figure 3 shows the segmentation results for one representative sample per modality across three mitral-valve imaging modalities: preoperative cardiac CT, intraoperative TEE, and surgical frame. Our model generates intact, smooth segmentation outlines, suppressing fragmented predictions and pixel-level artifacts, and enables reliable mitral-valve delineation under low contrast, acoustic interference and complex intraoperative conditions.

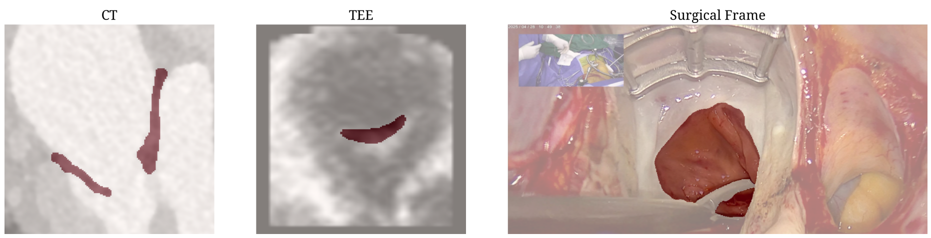

Fig. 3: Segmentation outputs of the proposed model on three mitral-valve modalities (CT, TEE, Surgical frame), yielding intact smooth contours free of jagged edges, missing boundaries and false-positive artifacts.

6 Conclusion

In this work, we propose a modality-adaptive hybrid-supervised framework for perioperative mitral valve segmentation across CT, TEE, and surgical video. By integrating UniMatch-based semi-supervision, multi-scale refinement, and boundary-aware losses, the proposed method alleviates noise, jagged edges, and incomplete contours, achieving robust performance across all MVAA subtasks. This decoupling is motivated by fundamental heterogeneity among different modalities. CT provides volumetric 3D representations, TEE suffers from acoustic artifacts, and surgical video is disturbed by instrument occlusion, which makes naive joint optimization vulnerable to negative transfer. Beyond the MVAA benchmark, the framework addresses a general challenge in multi-modal learning and is well suited to both preoperative planning and intraoperative guidance.

References

1. Garcia-Sayan E, Chen T, Khalique OK. Multimodality Cardiac Imaging for Procedural Planning and Guidance of Transcatheter Mitral Valve Replacement and Mitral Paravalvular Leak Closure. *Front Cardiovasc Med*. 2021 Feb 22.

2. Blanke P, Naoum C, Webb J, Dvir D, Hahn RT, Grayburn P, Moss RR, Reisman M, Piazza N, Leipsic J. Multimodality Imaging in the Context of Transcatheter Mitral Valve Replacement: Establishing Consensus Among Modalities and Disciplines. *JACC: Cardiovascular Imaging*, 2015, 8(10): 1191–1208.
3. Tan, B, Lin, X, Zhu, Q. et al. CT-Guided Preprocedural Assessment of TEER in Mitral Prolapse: When Calcification Limits TEE. J Am Coll Cardiol Case Rep. 2026 Jul, 31 (27) .
4. Saric M, Sadeghpour A, Alizade L, Alizadehasl A, Bertrand PB, Billick K, Chebrolu B, Faletra F, Kelsey A, Lang R, Levine RA, Pourafkari L, Walling S. Recommendations for the Identification and Mitigation of Cardiac Ultrasound Artifacts: A Guideline from the American Society of Echocardiography. *Journal of the American Society of Echocardiography*, 2026, 39(5).
5. Yamashita G, Nakano J, Sugaya A, Sakai J, Hirao S, Komiya T. Effectiveness of intraoperative endoscopic evaluation in aortic valve repair with valve-sparing aortic root replacement: a comparison of short- and mid-term results. *Interdisciplinary CardioVascular and Thoracic Surgery*, 2024, 38(5).
6. Cook V, Doyle M, Yan TD. Mini-access David Procedure with endoscopic assessment of aortic valve competency: the "Snorkelling" technique. *Annals of Cardiothoracic Surgery*. 2023;12(4):389-391.
7. Huynh PK, Kpodonu J, Le MHN, Nguyen D, Huynh P, Phan TQ, Olaniran OH, Tran T, Rutledge-Jukes H, Xu Y, Nguyen DH. 3-D TEE Mitral Valve Segmentation and Mesh Reconstruction with Real-Time Quality Assurance. In: *IEEE-EMBS International Conference on Biomedical and Health Informatics 2025*.
8. Bai, J., Khobo, I., Lu, Y., et al.: Landmark Detection Challenge for Intrapartum Ultrasound Measurement Meeting the Actual Clinical Assessment of Labor Progress. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. Zenodo (2025).
9. Zia, A., Berniker, M., Perreault, C., Nespolo, R., Jarc, A. Surgical Visual Understanding Challenge (SurgVU). EndoVis Grand Challenge, MICCAI 2025 (challenge description first posted on Zenodo, 2024). doi:10.5281/zenodo.14054184.
10. Chen M, Li X, Yang K, Taramasso M. AutoClip: AI-Guided TEE Semantic Segmentation for TEER A Proof-of-Concept Study. medRxiv. 2026.
11. He, Y., Guo, P., Tang, Y., Myronenko, A., Nath, V., Xu, Z., Yang, D., Zhao, C., Simon, B., Belue, M., Harmon, S., Turkbey, B., Xu, D., Li, W.: VISTA3D: A unified segmentation foundation model for 3D medical imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20863–20873 (2025).
12. M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, DINOv2: Learning Robust Visual Features without Supervision, *Transactions on Machine Learning Research*, Jan. 2024.
13. L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, Revisiting Weak-to-Strong Consistency in Semi-Supervised Semantic Segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 7236–7246.
14. A. Tarvainen and H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 1195–1204.

15. Hammad, M., Bai, J., Zhao, J., Jia, T., Zhang, X., Xu, X., Ma, J.& Li, S. (2026). Mitral Valve Anatomy Analysis Using Multimodal Imaging Data. Zenodo. International Conference on Medical Image Computing and Computer Assisted Intervention 2026 (MICCAI).
16. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, Cham, 2016, pp. 424–432.
17. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation, *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.
18. Bo Deng and Jieyun Bai, Baseline Method of the MVAA 2026 Challenge for Multimodal Mitral Valve Segmentation, in: The 1st MICCAI Workshop on Medical World Models, 2026.
19. MONAI Team. (2021). SegResNet. In *Medical Open Network for AI (MONAI)*.
20. Sihan Chen and Jianning Chi, Task-Specific Segmentation Frameworks for Multi-Modal Mitral Valve Tasks in MVAA 2026, in: The 1st MICCAI Workshop on Medical World Models, 2026.
21. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203–211.
22. Hui Xu, Yichen Niu, Hanlong Yin, Fangjin Guo and Yue Zhao, Task-Specific Segmentation Pipelines for the MVAA Challenge, in: The 1st MICCAI Workshop on Medical World Models, 2026.
23. Han Wu and Dinggang Shen. MVAA-UniSeg: A Query-Conditioned Universal Model with Modality-Specific Experts for Mitral Valve Segmentation. In: The 1st MICCAI Workshop on Medical World Models, 2026.