MVAA-UniSeg: A Query-Conditioned Universal Model with Modality-Specific Experts for Mitral Valve Segmentation

Han Wu[1,4] and Dinggang Shen[1,2,3(✉)]

[1] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China
[2] Shanghai United Imaging Intelligence Co. Ltd., Shanghai, China
[3] Shanghai Clinical Research and Trial Center, Shanghai, China
[4] Lingang Laboratory, Shanghai, China
dinggang.shen@gmail.com

Abstract. The mitral valve is a single anatomical structure, yet its appearance varies across imaging modalities, i.e., cardiac CT, 3D transesophageal echocardiography (TEE), and surgical video. Multi-modal mitral valve segmentation is thus usually solved by training one modality-specific *expert* model for each modality. Such *experts* are strong, but each of them has to learn from a single modality, and cannot combine information among all three modalities for better characterizing the same valve. The MICCAI 2026 MVAA challenge poses this problem in a setting much more challenging and closer to the clinical scenario, i.e., the three modalities differ greatly in the size of their labeled data along with a large amount of unlabeled data. We propose MVAA-UniSeg, a novel query-conditioned *universal model*, to segment the valve across modalities. Specifically, a per-task query is designed to condition a shared network at both its input and its bottleneck, while a semi-supervised branch is proposed to distill the unlabeled data into this unified representation. We further find that the two volumetric modalities, cardiac CT and TEE, benefit from being learned jointly, whereas adding the 2D video frames degrades all three tasks. We therefore share one network across CT and TEE, and keep a dedicated frame *expert* for video, trained with the unlabeled frames in a semi-supervised manner. Experiments on the held-out validation set show that MVAA-UniSeg outperforms both the CT and TEE *experts*. Our final solution, which combines it with the complementary CT *expert* and the frame *expert*, ranks first under the challenge's normalized score on both the validation and final test leaderboards. Our code will be released at `https://github.com/Fitz-Fitz/MVAA-UniSeg`.

Keywords: Mitral valve segmentation · Universal segmentation model · Cross-modal learning · Semi-supervised learning.

1 Introduction

Mitral valve segmentation is crucial for valvular heart disease analysis [9]. The appearance of the valve, however, varies greatly across imaging modalities. For

example, in cardiac CT, it is a thin, low-contrast leaflet embedded in a grayscale 3D volume. In 3D transesophageal echocardiography (TEE), it appears as two curved leaflet surfaces in a noisy ultrasound field [7]. In intra-operative surgical video frames, it is a highly reflective, partially occluded surface [21].

The standard approach for multi-modal mitral valve segmentation is to train a separate *expert* model for each modality. Each model sees only the images of its own modality, and is specially tuned for the modality's imaging physics and label set. nnU-Net [5,6] is the most representative example. Modality-specific *experts* set a strong baseline, and most automatic mitral valve methods are built in this way [1,2,13]. However, this approach ignores correlations among different modalities of the same valve, and requires one additional model for every new modality. Universal segmentation models can remove this redundancy by training a single network for multiple tasks and modalities. For example, DoDNet [18] and the CLIP-driven model [8] generate dynamic filters from a task encoding, and UniSeg [17] learns a universal task prompt to guide such a network across modalities. In these models, however, the task signal is injected late, at either the last layer [8,18] or the bottleneck [17], leaving the encoder unaware of the task being solved. They also assume that every task comes with a densely annotated dataset of comparable size, while clinical data rarely satisfy this assumption: most images carry no label [3,14].

The MICCAI 2026 MVAA (Mitral Valve Anatomy Analysis) challenge [4] poses multi-modal mitral valve segmentation in exactly this setting. It treats the mitral valve, imaged by cardiac CT, 3D TEE, and surgical video, as one *anatomical structure* captured through three *modalities*. Moreover, the supervision it provides is strongly unbalanced: CT has only 27 labeled volumes against 1040 unlabeled ones, TEE has 105 labeled volumes and no unlabeled pool, and video has 180 labeled frames against 1379 unlabeled ones. This raises two questions. First, how far does one shared representation extend across the three modalities? Second, can such a universal model use the abundant unlabeled data to match or even surpass the modality-specific *experts*?

We answer both questions with the following three contributions: 1) We propose MVAA-UniSeg, a novel query-conditioned universal model, to segment cardiac CT and 3D TEE within one network, outperforming both the CT and TEE *experts*. 2) MVAA-UniSeg is built on two designs: a per-task query first conditions the shared network at both its input (where it generates the kernel of a query-dynamic convolution) and its bottleneck (where it reads a content-adaptive prompt). Then, a semi-supervised branch is proposed to distill the unlabeled CT volumes into the same network, which improves segmentation on both modalities. 3) We measure how far this sharing extends, and find that adding the 2D frames to the shared network lowers accuracy on all three tasks. Hence, we combine MVAA-UniSeg with the CT *expert* and the frame *expert*, which ranks first under the challenge's normalized score on both the validation and final test leaderboards.

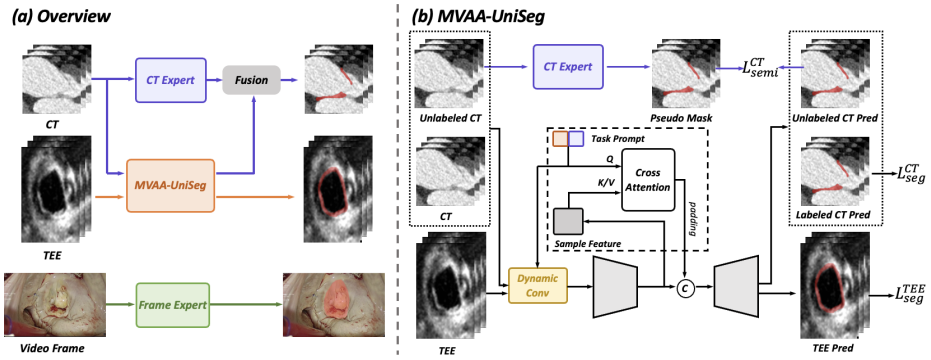

Fig. 1. (a) How the modality-specific *experts* and the cross-modal universal model MVAA-UniSeg work together for final prediction. **(b)** Overview of MVAA-UniSeg.

2 Method

2.1 Overview

Our final solution is shown in Fig. 1(a). It comprises several modality-specific *expert* models and one universal model (MVAA-UniSeg). We present MVAA-UniSeg in Sec. 2.2, the individual *expert* models in Sec. 2.3, and the final fusion strategy in Sec. 2.4.

2.2 MVAA-UniSeg: A Cross-Modal Universal Model for CT and TEE

Cardiac CT and 3D TEE are both volumetric views of the same valve, but the CT *expert* and the TEE *expert* share no parameters. We thus propose MVAA-UniSeg (Fig. 1(b)), a query-conditioned network, to segment both volumetric modalities.

Existing universal models differ mainly in where the task signal is injected, i.e., either at the last layer [8,18] or at the bottleneck [17]. In both cases, the encoder itself never sees the task signal. Instead, we inject the task signal twice. Specifically, the same per-task query is designed to condition the shared network at its input (through a query-dynamic convolution) and also at its bottleneck (through a content-adaptive prompt).

Query-dynamic convolution. The query-dynamic convolution makes the first convolution task-specific. MVAA-UniSeg couples a residual encoder E [6] and decoder D, shared by both modalities, with a pair of learnable task queries $\{\mathbf{q}_{\text{CT}}, \mathbf{q}_{\text{TEE}}\}$. Following the controller-based parameter generation of DoDNet [18], the query $\mathbf{q}_t$ of the current task $t \in \{\text{CT}, \text{TEE}\}$ parameterizes the first convolution of the shared stem S, i.e., a lightweight controller g maps $\mathbf{q}_t$ to the weight $\mathbf{W}_t$

and bias $\mathbf{b}_t$ of this convolution:

$$g(\mathbf{q}_t) = [\mathrm{vec}(\mathbf{W}_t), \mathbf{b}_t]\,, \qquad \tilde{x}_t = \mathrm{Conv}(x; \mathbf{W}_t, \mathbf{b}_t), \tag{1}$$

so the network knows its task from the first layer, and no separate stem per modality is needed.

Content-adaptive bottleneck prompt. The content-adaptive prompt injects the same task query at the bottleneck. The task-conditioned input passes through the shared encoder to a bottleneck feature $\mathbf{F} = E(S(x)) \in \mathbb{R}^{C\times d\times h\times w}$. Rather than selecting a fixed per-task map from a static prompt bank, the query $\mathbf{q}_t$ reads this feature by cross-attention to produce the prompt:

$$\mathbf{p}_t = \mathrm{CrossAttn}(\mathbf{q}_t, \mathbf{F})\,, \tag{2}$$

which is padded, concatenated back onto $\mathbf{F}$, and decoded by the shared decoder, $\hat{y}_t = D([\mathbf{F}, \mathbf{p}_t])$. We supervise both modalities on their labeled data by a per-task segmentation loss (Dice + cross-entropy, ℓ_{seg}):

$$\mathcal{L}_{\mathrm{seg}}^{\mathrm{CT}} = \mathbb{E}_{(x,y)\in\mathcal{D}_{\mathrm{CT}}^{\ell}}\, \ell_{\mathrm{seg}}(\hat{y}_{\mathrm{CT}}, y), \quad \mathcal{L}_{\mathrm{seg}}^{\mathrm{TEE}} = \mathbb{E}_{(x,y)\in\mathcal{D}_{\mathrm{TEE}}^{\ell}}\, \ell_{\mathrm{seg}}(\hat{y}_{\mathrm{TEE}}, y). \tag{3}$$

A semi-supervised CT branch. The semi-supervised branch exploits the unlabeled CT volumes. CT provides 27 labeled volumes, along with 1040 unlabeled volumes (which supervised training usually leaves unused). We distill the frozen CT *expert* of Sec. 2.3; specifically, it acts as a teacher to label every unlabeled volume with a soft pseudo-mask $T_{\mathrm{CT}}(x)$ through a distillation loss:

$$\mathcal{L}_{\mathrm{semi}}^{\mathrm{CT}} = \mathbb{E}_{x\in\mathcal{D}_{\mathrm{CT}}^{u}}\, \ell_{\mathrm{kd}}(\hat{y}_{\mathrm{CT}}, T_{\mathrm{CT}}(x))\,. \tag{4}$$

Thus, the CT branch becomes semi-supervised, where the 27 labeled volumes keep their supervised loss $\mathcal{L}_{\mathrm{seg}}^{\mathrm{CT}}$. TEE has no unlabeled pool, so we train it by supervised segmentation only.

Training objective. We train MVAA-UniSeg on both modalities jointly, by sampling them 1:1. The complete training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{seg}}^{\mathrm{CT}} + \mathcal{L}_{\mathrm{seg}}^{\mathrm{TEE}} + \mathcal{L}_{\mathrm{semi}}^{\mathrm{CT}}, \tag{5}$$

where $\mathcal{L}_{\mathrm{seg}}^{\mathrm{CT}}$ and $\mathcal{L}_{\mathrm{seg}}^{\mathrm{TEE}}$ are the per-task CT and TEE segmentation losses, and $\mathcal{L}_{\mathrm{semi}}^{\mathrm{CT}}$ is the CT distillation loss.

2.3 Modality-Specific Experts

We train one *expert* per modality, each seeing only the images of its specific modality. The three *experts* play different roles. The CT *expert* is a baseline, the teacher of MVAA-UniSeg's semi-supervised branch, and also a member of the CT fusion (Sec. 2.4). The TEE *expert* is a baseline only, and does not join the final solution. The frame *expert* alone produces the video masks, because surgical video falls outside the sharing boundary we measure in Sec. 3.4.

CT expert. For CT we use a 3D full-resolution nnU-Net [5] whose encoder and decoder are initialized from VolumeFusion (VolF) self-supervised pre-training [11] on the unlabeled volumes. VolF's pre-training, unlike encoder-only SSL, covers both encoder and decoder, and therefore narrows the gap between the pre-training and the downstream segmentation task. We finally fine-tune it on the 27 labeled cases.

TEE expert. For 3D TEE, we train a 3D full-resolution nnU-Net [5] with a residual encoder, using only the labeled volumes under full supervision.

Frame expert. For the 2D frames, we build several UNet++ [20] models whose encoders are initialized from DINOv3 [10] pretrained on the unlabeled frames. All models are then trained on the labeled frames for segmentation together with the 1379 unlabeled ones, in a semi-supervised manner by following UniMatch [16].

2.4 Final Prediction

CT. The CT mask combines MVAA-UniSeg and the CT *expert*. The two models predict the same target but make different errors, so we combine their segmentations, using voxelwise set union (the member set $\mathcal{M}$), followed by post-processing:

$$\hat{y}^{\mathrm{CT}} = \mathrm{PP}\big(\bigcup_{m\in\mathcal{M}} \mathrm{fg}(\hat{y}_m)\big)\,, \tag{6}$$

where $\mathrm{fg}(\cdot)$ is the foreground indicator, $\bigcup$ the voxelwise union, and PP largest-connected-component selection followed by hole filling. The union keeps the complementary segmentations of both models.

TEE. The TEE mask is produced by MVAA-UniSeg alone, followed by per-class largest-connected-component selection and hole filling.

Video Frame. The frame mask comes from the frame *expert* alone. We apply no post-processing. The connectivity filtering used in CT and TEE would only delete the thin, often-disconnected valve.

3 Experiments

3.1 Dataset and Evaluation Metrics

Dataset. The MVAA challenge [4] poses mitral valve segmentation across three modalities, each with its own label set. *Task 1 (cardiac CT)* provides 27 labeled and 1040 unlabeled volumes together with a 30-case held-out set. The target is the binary mitral valve. *Task 2 (3D TEE)* provides 105 labeled volumes with two leaflet classes (anterior and posterior, scored per class) and a 20-case held-out set. *Task 3 (surgical video frames)* provides 180 annotated frames (118 containing the

valve) and 1379 unlabeled frames, with a 48-frame held-out set. These held-out sets are used in the validation phase, while a separate hidden set (whose size is not released) is used in the final-test phase. No ground truth is released for either phase, so all held-out scores we report come from the challenge's CodaBench evaluation server.

Evaluation metrics. Following the challenge protocol [4], each task is measured by the Dice similarity coefficient (DSC), the Hausdorff distance (HD), and the average surface distance (ASD). Teams are ranked independently on every metric and each rank is converted into normalized points, i.e., 100 for the first and 60 for the tenth with linear interpolation, $p(r) = 100 - \frac{40}{9}(r-1)$. Because Task 2 is built on a public dataset, every eligible team receives 100 points on its three metrics. The overall score is the mean of these nine values. For reference, we also report the sum of the ranks that actually decide this score.

3.2 Implementation Details

The volumetric models (the CT and TEE *experts* and MVAA-UniSeg) are built on nnU-Net v2.7 [5,6] (SGD, momentum 0.99, poly-decayed LR 10^{-2}, deep supervision). The CT *expert* is a plain 3D nnU-Net (patch $112 \times 128 \times 160$, 250 epochs) initialized by VolF-based self-supervised pre-training on the unlabeled CT pool [11]. The TEE *expert*, used here only as a baseline, is a full-resolution ResEnc-M nnU-Net trained under full supervision on the labeled TEE volumes. MVAA-UniSeg uses a shared ResEnc-M network at a $112 \times 128 \times 160$ patch, in which a 64-dim per-task query directly parameterizes the first stem convolution through a lightweight controller. We train it from scratch on CT+TEE for 500 epochs with CT-branch distillation (1:1 sampling, 50-epoch ramp-up). We train the frame *expert* at 512×896 resolution with AdamW for 140 epochs. At inference, we ensemble the folds by softmax averaging (5 for CT/TEE, 6 leave-one-recording-out for frames), with largest-connected-component and hole-filling post-processing for CT/TEE and none for frames. Training uses NVIDIA A100-80GB.

3.3 Comparison with Challenge Participants

On the public validation leaderboard, our solution ranks first among 56 teams with **96.54**, ahead of wyss520 (90.62) and mgazda (89.63) (Tab. 1a). Among the ten teams listed, it obtains the best CT DSC and HD, the best TEE HD and ASD, and the best frame DSC, and is second on frame HD and ASD. The official ranking is decided on the hidden final-test phase, where our solution scores **94.07** and ranks first among 37 teams (Tab. 1b), leading the runner-up wyss520 (92.59) by 1.48 points and the third-ranked mgazda (90.12) by 3.95 points. Its rank sum is 18, against 21 and 26 for the two runners-up. The lead comes from CT and video together rather than one metric. Among the top ten teams listed, our entry achieves the best CT DSC and ASD, and is second on frame DSC and HD. wyss520 takes the best frame DSC and ASD, but, on CT alone, it falls nine

Table 1. Top ten under the normalized score on **(a)** the validation and **(b)** the final test leaderboard, both phases closed. RankSum is the sum of the ranks that decide the score; lower is better. **Best** and second-best per metric among the listed teams.

#	Team snapshot of 2026-08-07	Score	Rank Sum	Task 1 (CT)			Task 2 (TEE)			Task 3 (frame)		
				DSC	HD	ASD	DSC	HD	ASD	DSC	HD	ASD
(a) Validation phase (comp. 15662, 56 teams):												
1	**Fitz (ours)**	**96.54**	**13**	**86.67**	**4.15**	0.256	86.22	**9.07**	**0.479**	**83.23**	65.44	10.685
2	wyss520	90.62	25	86.42	4.36	0.267	**86.51**	9.15	0.518	83.17	**63.57**	**10.139**
3	mgazda	89.63	27	86.53	4.25	**0.246**	84.50	10.20	0.603	81.66	68.15	11.706
4	ne_gi_chi__	85.68	35	86.34	4.42	0.255	86.26	9.59	0.526	82.29	67.91	11.181
5	willenhou	79.26	48	86.11	4.36	0.268	85.20	9.92	0.566	81.64	67.68	11.169
6	jguoaz	78.77	49	86.17	4.39	0.265	85.75	9.69	0.542	82.04	71.38	11.082
7	xiaoji	77.78	51	86.36	4.36	0.264	85.18	9.80	0.550	81.55	68.99	11.969
8	cjunxiao	72.35	62	86.29	4.46	0.253	85.42	9.59	0.555	81.70	71.99	12.221
9	xuhui	61.48	84	86.19	4.26	0.250	85.37	10.96	0.579	81.21	95.19	13.691
10	yixinchen0320	59.51	88	86.07	4.60	0.272	85.94	9.67	0.513	81.53	71.77	11.941
(b) Final test phase (comp. 17301, 37 teams):												
1	**Fitz (ours)**	**94.07**	**18**	**85.36**	5.07	**0.341**	90.02	6.15	0.295	85.01	162.09	36.139
2	wyss520	92.59	21	84.62	5.10	0.444	86.08	8.97	0.529	**85.46**	167.57	**28.481**
3	mgazda	90.12	26	84.94	4.96	0.404	85.83	9.17	0.497	83.86	170.50	35.771
4	liyilin	77.78	51	83.90	5.95	0.554	85.69	10.41	0.512	84.60	**146.30**	31.177
5	hars25	74.32	58	84.43	**4.94**	0.416	**94.64**	**3.15**	**0.137**	81.51	243.28	105.187
6	s2119	73.83	59	83.89	5.83	0.519	85.74	9.32	0.535	83.34	173.07	36.512
7	lizheng167	72.35	62	82.66	5.25	0.360	81.41	11.76	0.769	84.69	181.59	60.443
8	lyuyangtong	69.88	67	84.15	6.45	0.554	78.18	17.50	1.070	81.74	177.77	34.686
9	xuhui	68.89	69	84.08	5.33	0.478	85.28	10.43	0.546	80.13	278.82	61.115
10	cjunxiao	65.93	75	83.85	6.28	0.506	85.22	9.04	0.563	80.70	184.46	38.447

ranks behind our entry. The two phases score on two different hidden sets, so their values are not comparable across two phases. That is, every team listed in both parts of Tab. 1 could have larger frame HD and ASD in the final-test phase. Teams are ranked within each phase, so this shift does not affect the final ranking.

3.4 Ablation Studies

Modality-specific experts. We first tune one *expert* for each modality. The baselines in Tab. 2 are a plain nnU-Net for CT, a fully-supervised nnU-Net for TEE, and UniMatch [16] for frames. Tuning raises CT DSC from 85.80 to 86.21 with VolF-based pre-training, TEE DSC from 84.75 to 85.19 with a residual encoder, and frame DSC from 82.06 to 82.90 with DINOv3-based pre-training.

Separate learning for surgical video (2D frames). When all three tasks are learned in the same network, MVAA-UniSeg reaches only 53.17 DSC on the surgical frames, far below the frame *expert* (82.90), and its HD and ASD are dominated by the large penalty the challenge assigns to frames (on which no valid segmentations are predicted). Including the frames also lowers CT and TEE DSC, from 86.41/86.22 to 86.21/85.33 (Tab. 2). As is evident, the 2D frames are very different from the two volumetric modalities. Forcing all three into a single network can cause negative transfer on every task. We thus apply universal learning only to CT and TEE, while segmenting video with the frame *expert*.

Table 2. Ablation on the validation leaderboard across all nine metrics (DSC, HD, ASD per task). ⋆: final solution; **best** in bold, and second-best underlined.

Configuration	T1 (CT)			T2 (TEE)			T3 (frame)		
	DSC	HD	ASD	DSC	HD	ASD	DSC	HD	ASD
Modality-specific experts:									
Baseline (nnU-Net / nnU-Net / UniMatch [16])	85.80	4.63	0.28	84.75	10.17	0.60	82.06	73.01	12.19
Tuned (VolF / ResEnc-M / DINOv3)	86.21	4.63	0.27	85.19	10.20	0.56	82.90	65.86	10.80
Other universal models:									
DoDNet [18]	86.26	4.54	0.27	85.63	9.72	0.54	–	–	–
UniSeg [17]	86.25	4.56	0.27	85.75	9.35	0.53	–	–	–
MVAA-UniSeg:									
all three tasks in one net (CT + TEE + Frame)	86.21	4.47	0.26	85.33	10.29	0.55	53.17	187585.11	187515.54
MVAA-UniSeg (CT+TEE)	86.41	4.33	0.26	**86.22**	**9.07**	**0.48**	–	–	–
MVAA-UniSeg ablation (CT + TEE):									
w/o semi branch	86.27	4.60	0.26	86.03	9.24	0.50	–	–	–
w/o query-dynamic conv	86.33	4.46	0.26	86.09	9.29	0.52	–	–	–
Final submission:									
⋆ MVAA-UniSeg + CT *expert* + frame *expert*	**86.67**	**4.15**	**0.26**	**86.22**	**9.07**	**0.48**	**83.23**	**65.44**	**10.69**

Universal learning across CT and TEE. Every universal configuration outperforms the tuned CT and TEE *experts* (86.21/85.19 DSC): DoDNet and UniSeg reach 86.26/85.63 and 86.25/85.75, while MVAA-UniSeg is best on both tasks at 86.41/86.22 (Tab. 2). Removing the semi-supervised branch reduces DSC by 0.14/0.19 to 86.27/86.03 and worsens HD from 4.33/9.07 to 4.60/9.24. Although the branch distills unlabeled CT only, TEE also improves, since the two modalities meet only inside the shared network. Removing the query-dynamic convolution reduces DSC to 86.33/86.09 and worsens HD from 4.33/9.07 to 4.46/9.29, which indicates that early task conditioning complements the bottleneck prompt.

4 Conclusion

In this paper, we explored two questions: (1) how far one shared representation can extend across the three modalities of the MVAA challenge, and (2) whether a universal model can use the unlabeled data to surpass the modality-specific *experts*. Accordingly, we proposed MVAA-UniSeg, to condition a shared network with a per-task query at both its input and its bottleneck, and further distill the unlabeled data. Trained jointly on CT and TEE, MVAA-UniSeg outperforms both modality-specific *experts*, and also improves TEE's segmentation, even though only CT contributes unlabeled data. But, for surgical video (2D frames), we still use its own modality-specific *expert*. Thus, our final solution combines MVAA-UniSeg with the CT *expert* and the frame *expert*, and ranks first on both the validation and final test leaderboards.

Note that MVAA-UniSeg is currently trained on *unpaired* data from *different* patients. Paired CT, TEE, and video from the same patient would enable direct cross-modal learning, and could bring the video frames into a joint shared network (with CT and TEE). Also, such paired data can encourage multi-modal alignment [15,19] and agent-based systems for complex multi-modal diagnosis [12].

References

1. Andreassen, B.S., Veronesi, F., Gerard, O., Solberg, A.H.S., Samset, E.: Mitral annulus segmentation using deep learning in 3-D transesophageal echocardiography. IEEE Journal of Biomedical and Health Informatics **24**(4), 994–1003 (2019)
2. Carnahan, P., Moore, J., Bainbridge, D., Eskandari, M., Chen, E.C., Peters, T.M.: DeepMitral: Fully automatic 3D echocardiography segmentation for patient specific mitral valve modelling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 459–468. Springer (2021)
3. Chen, H., Li, Y., Yang, L., Wu, H., Zhou, L., Sun, K., Shen, D.: A semi-supervised knowledge distillation framework for left ventricle segmentation and landmark detection in echocardiograms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 34–43. Springer (2025)
4. Hammad, M., Bai, J., Zhao, J., Jia, T., Zhang, X., Xu, X., Ma, J., Li, S.: Mitral valve anatomy analysis using multimodal imaging data. MICCAI 2026 Challenge, `https://doi.org/10.5281/zenodo.19726755` (2026)
5. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
6. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 488–498. Springer (2024)
7. Lang, R.M., Badano, L.P., Tsang, W., Adams, D.H., Agricola, E., Buck, T., Faletra, F.F., Franke, A., Hung, J., de Isla, L.P., et al.: EAE/ASE recommendations for image acquisition and display using three-dimensional echocardiography. European Heart Journal–Cardiovascular Imaging **13**(1), 1–46 (2012)
8. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., Landman, B.A., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: CLIP-driven universal model for organ segmentation and tumor detection. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21095–21107. IEEE (2023)
9. Nkomo, V.T., Gardin, J.M., Skelton, T.N., Gottdiener, J.S., Scott, C.G., Enriquez-Sarano, M.: Burden of valvular heart diseases: a population-based study. The Lancet **368**(9540), 1005–1011 (2006)
10. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)
11. Wang, G., Fu, J., Wu, J., Luo, X., Zhou, Y., Liu, X., Li, K., Lin, J., Shen, B., Zhang, S.: Volume fusion-based self-supervised pretraining for 3D medical image segmentation. IEEE Transactions on Image Processing (2025)
12. Wang, S., Zhao, Z., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Interactive computer-aided diagnosis on medical image using large language models. Communications Engineering **3**(1), 133 (2024)
13. Wu, H., Chen, H., Zhou, L., Xu, Q., Cui, Z., Shen, D.: Semi-supervised landmark tracking in echocardiography video via spatial-temporal co-training and perception-aware attention. IEEE Transactions on Medical Imaging (2026)
14. Wu, H., Wang, C., Cui, Z.: Dual cross-image semantic consistency with self-aware pseudo labeling for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging (2025)

15. Xiong, X., Jiang, C., Wu, H., Zhang, X., Wu, P., Mi, J., Song, Y., Zhang, X., Zhang, J., Wu, D., et al.: Two-stage robust 3D CTA–2D DSA alignment via vascular-aware rigid and pyramid-based hierarchical non-rigid registration. Medical Image Analysis p. 104150 (2026)
16. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7236–7246. IEEE (2023)
17. Ye, Y., Xie, Y., Zhang, J., Chen, Z., Xia, Y.: UniSeg: A prompt-driven universal segmentation model as well as a strong representation learner. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 508–518. Springer (2023)
18. Zhang, J., Xie, Y., Xia, Y., Shen, C.: DoDNet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1195–1204. IEEE (2021)
19. Zhao, Z., Liu, Y., Wu, H., Wang, M., Li, Y., Wang, S., Teng, L., Liu, D., Cui, Z., Wang, Q., et al.: CLIP in medical imaging: A survey. Medical Image Analysis **102**, 103551 (2025)
20. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Transactions on Medical Imaging **39**(6), 1856–1867 (2019)
21. Zia, A., Berniker, M., Perreault, C., Nespolo, R., Jarc, A.: Surgical visual understanding challenge (SurgVU). MICCAI Challenge, `https://doi.org/10.5281/zenodo.14054184` (2024)