A Multimodal Segmentation Pipeline for Mitral Valve Anatomy Analysis in Cardiac CT, 3D Transesophageal Echocardiography, and Surgical Video

Sanket Kachole[1,2] and Spyridon Bakas[1,2,3,4]

[1] Division of Computational Pathology, Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, IN, USA
[2] IU Melvin and Bren Simon Comprehensive Cancer Center, Indianapolis, IN, USA
[3] Departments of Biostatistics and Health Data Science; Radiology and Imaging Sciences; Neurological Surgery; Indiana University School of Medicine, Indianapolis, IN, USA
[4] Department of Computer Science, Luddy School of Informatics, Computing, and Engineering, Indiana University, Indianapolis, IN, USA
spbakas@iu.edu

Abstract. Mitral valve repair spans cardiac computed tomography (CT) planning, intraoperative three-dimensional transesophageal echocardiography (TEE) guidance, and surgical video, motivating a unified multimodal segmentation framework. We combine self-configuring three-dimensional networks for CT and TEE with a mean-teacher ensemble of ImageNet-pretrained two-dimensional video networks, label auditing and cleaning, connected-component filtering, and empty-mask rescue; development used the MVAA 2026 dataset comprising 27 labelled and 1,040 unlabelled CT volumes, 105 labelled TEE volumes, and 180 labelled and 1,379 unlabelled video frames. On the held-out test set, Dice/HD/ASD were 0.839/5.43 mm/0.43 mm for CT, 0.853/9.19 mm/0.56 mm for TEE, and 0.755/236.29/63.76 pixels for video; the principal gains came from label correction and empty-mask rescue rather than architectural changes, reducing video HD from 520.44 to 236.29 pixels. Code and trained weights are available on GitHub and Zenodo.

Keywords: Mitral valve · Multimodal segmentation · Cardiac CT · Transesophageal echocardiography · Surgical video

1 Introduction

Valvular heart disease is common in older adults, with mitral regurgitation among its most prevalent forms; population studies guide clinical and technical priorities, particularly where early detection improves outcomes [22,24]. Mitral repair spans preoperative computed tomography (CT), intraoperative three-dimensional transesophageal echocardiography (3D TEE), and direct surgical video, yet manual valve delineation is slow and observer-dependent. Automatic methods have progressed from atlas- and model-based approaches [7,23]

to encoder–decoder networks [26] and 3D TEE convolutional networks [1,2,20], while annotated benchmarks have advanced video and region-based analysis across monitoring applications [6,13]. Nevertheless, most cardiac segmentation studies remain modality-specific [5], despite evidence that multisource medical-imaging models can improve diagnosis and prognosis [16,29]. The MVAA 2026 challenge addresses this gap through a unified three-task benchmark covering CT, 3D TEE, and surgical video with common overlap and distance metrics.

We present a unified challenge framework using residual-encoder nnU-Net models for both volumetric tasks, trained by five-fold cross-validation and deployed as five-model ensembles [8,9]. For video, six UNet++ models with ImageNet-pretrained EfficientNet-B4 encoders were trained using different seeds and video-level holdouts, incorporating unlabelled frames through mean-teacher learning, and their probability maps were averaged [30,27,28]. Label auditing identified tiny isolated video regions, which were removed from both training labels and predictions; for CT, fold-wise pseudo-labels were retained only when the folds agreed closely [18]. A single container generated predictions for all three tasks in one run.

2 Materials and Methods

2.1 Dataset

Table 1 summarises cardiac CT with one mitral-valve foreground class, 3D TEE with two unnamed leaflet classes (labels 1 and 2), and surgical-video frames with one foreground class. Connected-component auditing found that all 27 CT labels contained a single region. TEE class 1 comprised a single region in 102 of 105 volumes, whereas class 2 did so in only 72, reached 11 regions in one case, and had a mean maximum stray-voxel distance of 1.58 mm versus 0.02 mm. Video labels were noisiest, averaging 23 regions per frame (maximum 107), with a median region size of two pixels and only 21% single-region frames; removing regions smaller than 50 pixels discarded only 0.10% of the labelled area while reducing the average count from 23 to 3, confirming that most were speckle. Two properties of Task 3 shaped our design. First, 62 of the 180 labelled frames contain no valve, and one further frame was withdrawn by the organisers, leaving 117 for supervised training. Second, every official validation frame contains the valve, so the validation score cannot measure how a model behaves when the valve is absent. Our training split is made at the level of whole videos rather than individual frames: every frame carries its source video identifier, and each video is assigned entirely to training or entirely to internal validation, so no frame from a held-out video is ever seen during training. Each of the six models holds out a different pair of videos, drawn by a distinct random seed. The 62 valve-free frames and the unlabelled frames are identified separately by whether any foreground is present, and are handled accordingly in the split.

2.2 Method

Figure 1 shows the pipeline from input to output. The three tasks share the same sequence of stages, namely preprocessing and label auditing, training of a set of networks, ensembled inference with test-time augmentation, and a post-processing step, but they differ in the network family and in the post-processing rule.

Preprocessing and label auditing. For Tasks 1 and 2, nnU-Net automatically resampled each volume to the dataset's median spacing and normalised intensities; the selected spacings matched our manual estimates. For Task 3, we excluded the withdrawn and valve-free frames, resized the remainder to 448×800 pixels, applied ImageNet normalisation, and removed training-label regions smaller than 100 pixels. Cleaning primarily improved distance metrics because speckled labels induce speckled predictions, and even one distant pixel can greatly inflate the Hausdorff distance [25].

Volumetric segmentation for Tasks 1 and 2. Both volumetric tasks used the large-preset nnU-Net residual-encoder configuration, trained for 250 epochs with the framework's default loss and optimiser [9]. nnU-Net selected the patch size, batch size, and augmentation policy from each dataset's properties, consistent with evidence that sensor-specific augmentation outperforms generic augmentation [21]. Instance normalisation accommodated the small batch sizes. We retained all five models from the standard cross-validation split, avoiding separate architecture and schedule tuning for modalities with markedly different voxel spacings and labelled sample sizes.

Video segmentation for Task 3. For surgical video, we used a UNet++ decoder with an ImageNet initialised EfficientNet-B4 encoder; with only 117 labelled frames, pretraining mattered more than architecture, unlike data rich settings where architectural or input representation changes can yield large gains [30,27,10,15,11]. Models were trained for up to 200 epochs with early stopping after 60 epochs without improvement, using Dice and focal losses weighted

Table 1. Summary of the challenge data. Counts are volumes for Tasks 1 and 2 and individual frames for Task 3. Spacings are median values per axis.

	Task 1 (CT)	Task 2 (3D TEE)	Task 3 (video)
Image type	3D volume	3D volume	2D RGB frame
Spacing (mm) / size (px)	0.36/0.36/0.50	0.37/0.54/0.23	1280×720
Labelled (used to train)	27 (27)	105 (105)	180 (117)
Unlabelled	1,040	0	1,379
Validation cases	30	20	48
Foreground classes	1	2	1

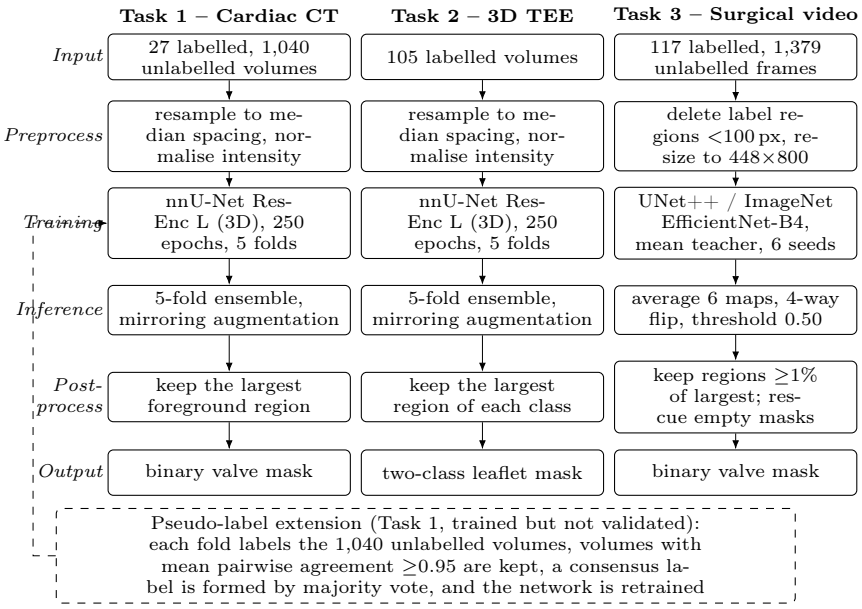

Fig. 1. The pipeline from input to output. The three tasks share the same sequence of stages but use different networks and different post-processing rules. The dashed box shows the pseudo-label extension built for the CT task, which was trained but could not be validated before the deadline.

0.7 and 0.3, with $\gamma = 2.0$ and $\alpha = 0.75$. AdamW used an initial learning rate of 2×10^{-4}, five epochs of linear warmup, cosine decay to 1×10^{-6}, weight decay of 1×10^{-5}, gradient clipping at an L2 norm of 1.0, and mixed precision. Each batch contained four labelled and four unlabelled frames, with labelled sampling weighted by foreground content using power 0.5 and clipping to $[0.5, 4.0]$. Mean teacher training used an exponential moving average decay of 0.99, 20 supervised epochs, and a 30 epoch ramp to an unsupervised weight of 0.6; probabilities above 0.7 and below 0.1 defined positive and negative targets, and pseudo labels required at least 80 pixels [28]. Six models with seeds 42 to 47 each held out a different video pair, providing diversity in both initialisation and training data.

Inference and post-processing. For Tasks 1 and 2 the five models are ensembled with mirroring test-time augmentation, following the default of the framework; for Task 3 each of the six models is applied with four-way flip augmentation, the six probability maps are averaged and the mask is taken at a threshold of 0.50. Combining predictions late, at the level of output probabilities, is a simple and robust alternative to fusing information inside the network [12]. Post-processing is then applied per task: the largest foreground region is kept for Task 1, the largest region of each class is kept for Task 2, and for Task 3 every region smaller than 1% of the largest region is deleted. Task 3 adds one further step: if

Table 2. Computational cost of the pipeline. Training times are wall-clock; the Task 3 figure is per model and Tasks 1 and 2 are per fold. Inference is the complete three-task run on the evaluation server.

	Task 1 (CT)	Task 2 (TEE)	Task 3 (video)	Full pipeline
Backbone	nnU-Net ResEnc L	nnU-Net ResEnc L	UNet++/EffNet-B4	–
Members	5 folds	5 folds	6 seeds	16 models
Epochs	250	250	200 (early stop)	–
Train time/member	4 h	4 h	27 min	–
GPU memory	12 GB	12 GB	10 GB	–
Inference time	–	–	–	3770 s (1 V100)
Model size	–	–	–	4.6GB (from 10.1GB)

the thresholded mask for a frame is empty, an unconditional rescue selects the highest-probability pixels, namely the greater of 64 pixels or 0.5% of the frame. The rescued mask is then passed through the same connected-component rule, retaining only components at least 1% of the largest.

Pseudo-label extension for Task 1. The CT dataset included 1,040 unlabelled volumes, nearly 40 times the 27 labelled cases. Each fold model independently segmented these volumes, and the mean pairwise Dice across the five predictions measured confidence. Agreement was high, with a mean of 0.9264, a median of 0.9474, and no empty predictions. In total, 934 volumes reached 0.90 and 432 reached 0.95. Predictions above each threshold were combined by majority vote and added to the labelled set for retraining. However, the folds shared approximately 78% of their training data, which likely overstated confidence and made the 0.95 threshold safer. Generative synthesis offers another response to label scarcity but was not pursued [4,3,14].

Implementation details. The pipeline was implemented in PyTorch 2.5.1 with CUDA 12.1, using nnU-Net version 2.8.1 for Tasks 1 and 2 and the segmentation-models-pytorch library for Task 3. Training used one NVIDIA RTX 6000 Ada card with 48 GB of memory and, for parallel folds, NVIDIA V100 cards with 32 GB. On the evaluation server, using a single V100 card, the complete run took 3,770 seconds against a limit of 21,600 seconds, and reducing the checkpoints to inference weights lowered the model files from 10.1 GB to 4.6 GB.

3 Results

The challenge reports the Dice similarity coefficient (DSC), the maximum Hausdorff distance (HD) and the average surface distance (ASD) [19]. Distances use physical spacing where it is available, so Tasks 1 and 2 are scored in millimetres and Task 3, whose frames carry no spacing, in pixels. Table 3 gives the development history on the official validation set and Table 4 gives the final held-out test scores. One caveat applies throughout: the organisers corrected the Task 3

Table 3. Official validation scores. Higher is better for DSC, lower for HD and ASD. Distances are millimetres for Tasks 1 and 2 and pixels for Task 3. The configuration carried forward is shown in bold. ASD was not recorded for the 2D nnU-Net run.

Task	Configuration	DSC	HD	ASD
1 (CT)	provided 3D U-Net baseline	0.8035	8.020	0.4230
1 (CT)	nnU-Net ResEnc L, 5-fold ensemble	0.8575	4.365	0.2715
1 (CT)	+ automatic post-processing	**0.8582**	**4.225**	**0.2612**
2 (TEE)	provided 3D U-Net baseline	0.7232	29.430	2.2700
2 (TEE)	nnU-Net ResEnc L, single fold	0.8450	11.290	0.6570
2 (TEE)	nnU-Net ResEnc L, 5-fold ensemble	0.8514	10.860	0.6094
2 (TEE)	+ automatic post-processing	**0.8522**	**10.169**	**0.5853**
3 (video)	2D nnU-Net, original labels	0.7304	132.710	–
3 (video)	1 UNet++, original labels, 20% filter	0.7956	85.880	14.5500
3 (video)	6 UNet++, original labels, no filter	0.8058	161.700	18.1100
3 (video)	6 UNet++, original labels, 20% filter	0.8061	83.880	14.1600
3 (video)	6 UNet++, cleaned labels, no filter	0.8085	97.460	13.4600
3 (video)	6 UNet++, cleaned labels, 1% filter	**0.8089**	**78.570**	**13.0700**

HD and ASD rule late in the validation phase, so the Task 3 distances in Table 3 were produced under the previous rule and are not directly comparable with those in Table 4.

Tasks 1 and 2. Replacing the provided baseline with the nnU-Net residual encoder configuration increased CT Dice from 0.8035 to 0.8575 and reduced Hausdorff distance from 8.02 to 4.37 mm. For echocardiography, a single fold reached 0.8450 compared with the baseline score of 0.7232, although part of this difference resulted from an unstable baseline batch size. Ensembling all five folds added 0.006 Dice, while automatic postprocessing improved all six metrics, most notably reducing echocardiography Hausdorff distance from 10.86 to 10.17 mm. Training for 500 rather than 250 epochs changed cross validation Dice by only 0.0016 for CT and 0.0001 for echocardiography. A ten model ensemble combining both schedules increased echocardiography Dice by only 0.0009 while worsening Hausdorff distance from 10.169 to 10.642 mm, so both variants were rejected.

Task 3. The single most useful choice for the video task was to initialise the encoder from ImageNet rather than from random weights, which raised the internal validation Dice from about 0.58 to about 0.79; the second was to treat label noise directly. Cleaning the training labels reduced the unfiltered Hausdorff distance of the six-model ensemble from 161.70 to 97.46 pixels and the average surface distance from 18.11 to 13.46, and deleting predicted regions smaller than 1% of the largest then reduced the Hausdorff distance further to 78.57. Across the filter sweep the Hausdorff distance was 78.57, 79.16, 80.11, 80.89, 81.08, 84.40 and 84.40 pixels at thresholds of 1, 2, 3, 5, 7, 12 and 20%, so the mildest filter

Table 4. Held-out test scores. Tasks 1 and 2 were identical across all four submissions because their code paths were unchanged. Variants differ only in how Task 3 handles a frame whose thresholded mask is empty. Variant v2 was submitted.

Task	Variant	DSC	HD	ASD
1 (CT)	all submissions	0.8394	5.426	0.4327
2 (TEE)	all submissions	0.8530	9.190	0.5649
3 (video)	v1, no empty-mask handling	0.7502	520.441	386.683
3 (video)	v3, rescue gated at $p > 0.50$	0.7502	520.441	386.683
3 (video)	v4, rescue gated at $p > 0.35$	0.7502	520.441	386.683
3 (video)	**v2, unconditional rescue**	**0.7551**	**236.294**	**63.763**

was the best. Ensembling six models rather than one added 0.011 Dice, whereas a twelve-model ensemble mixing cleaned and original labels gave a similar Dice of 0.8087 but a worse Hausdorff distance of 85.42. Most of these comparisons hold the training budget constant and vary a single factor: ensembling is isolated by comparing one and six models under the same labels and filter, label cleaning by comparing original and cleaned labels under the same six-model ensemble and no filter, and filter strength by the sweep over a single set of six cleaned models. The one comparison that is not compute-matched is between the 2D nnU-Net and the UNet++ models, which differ in architecture and in pretraining at once, since the nnU-Net is trained from scratch while the UNet++ encoder is initialised from ImageNet; that difference should therefore be read as the combined effect of architecture and pretraining rather than of architecture alone.

Failure analysis and negative results. The first Task 3 test submission achieved Dice 0.7502 but a Hausdorff distance of 520.44 pixels because the ensemble produced empty masks for references containing the valve, causing catastrophic case-level penalties rather than broadly poorer delineation [25]. Unconditional rescue (v2 in Table 4) reduced Hausdorff distance to 236.29 pixels and average surface distance from 386.68 to 63.76, while increasing Dice to 0.7551, confirming that the affected frames contained the valve. The gated variants failed: v3 cannot activate at 0.50 when the mask thresholded at 0.50 is empty, while v4 produced scores identical to no rescue at 0.35, showing that the failures had maximum probabilities below 0.35; only unconditional v2 recovered them.

Before rescue, all 62 labelled valve-free frames had zero predicted area. However, unconditional rescue would make these masks non-empty and could therefore introduce false positive predictions when the valve is absent. This limitation is not captured by the official validation set, in which every frame contains the valve. Four other changes gave no improvement: higher input resolution, threshold sweeps between 0.10 and 0.70, boundary loss (-0.017 Dice and $+2.04$ HD) [17], and a 2D nnU-Net, which achieved 0.7926 in cross validation but 0.7304 officially. The large reduction in distance errors with only a small Dice increase

suggests that the rescue primarily corrected a few catastrophic empty predictions rather than broadly changing performance. The exact number of rescued test frames and their case-level metric distribution cannot be reported because the evaluation server returned only aggregate scores and the test frames are held by the organisers.

Pseudo-labels for Task 1. Agreement filtering produced two candidate training sets, one of 961 volumes at a threshold of 0.90 and one of 459 at 0.95, each being the volumes passing that threshold together with the 27 labelled cases, and two five-fold training runs (one per candidate set, ten fold models in total) ran to completion. Their reported validation Dice of 0.96 to 0.97 cannot be used, because the cross-validation split covers the pseudo-labelled cases the models themselves generated, so the models are scored partly against their own output. These pseudo-label experiments are therefore exploratory only and are reported as an invalidated negative result: they were never submitted, and the submitted CT model uses the 27 labelled volumes alone.

4 Conclusion

We developed a unified framework for mitral-valve segmentation across cardiac CT, 3D TEE, and surgical video, combining residual-encoder three-dimensional networks with a mean-teacher ensemble of transfer-learned two-dimensional video networks and task-specific post-processing. Held-out test Dice scores were 0.839, 0.853, and 0.755, with end-to-end inference within the challenge time limit; label auditing and empty-prediction handling mattered more than architecture on these small, noisy datasets, while ImageNet pretraining outweighed architecture search using only 117 labelled frames. The video model nevertheless missed valves in some unseen frames (maximum probability < 0.35), meaning the fallback addresses the symptom rather than the cause; moreover, CT pseudo-labeling was trained but not validated. Qualitative prediction examples for all three modalities can be produced from the released code and are not reproduced here for space.

Acknowledgments. This work was partially supported by the National Cancer Institute, National Institutes of Health, through the Informatics Technology for Cancer Research program (U24CA279629), and by Lilly Endowment, Inc., through the Indiana University Pervasive Technology Institute. The content is solely the authors' responsibility and does not represent the official views of the NIH or any other funding body.

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

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. Chandrasekaran, M., Kachole, S., Francik, J., Makris, D.: PGcGAN: Pathological gait-conditioned gan for human gait synthesis. arXiv preprint arXiv:2603.14409 (2026)
4. Chandrasekaran, M., Kachole, S., Francik, J., Makris, D.: LLM-conditioned synthesis of pathological gaits via structured gait-language representations. arXiv preprint arXiv:2606.06048 (2026)
5. Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., Rueckert, D.: Deep learning for cardiac image segmentation: a review. Frontiers in cardiovascular medicine **7**, 25 (2020)
6. Huang, X., Kachole, S., Ayyad, A., Naeini, F.B., Makris, D., Zweiri, Y.: A neuromorphic dataset for tabletop object segmentation in indoor cluttered environment. Scientific data **11**(1), 127 (2024)
7. Ionasec, R.I., Voigt, I., Georgescu, B., Wang, Y., Houle, H., Vega-Higuera, F., Navab, N., Comaniciu, D.: Patient-specific modeling and quantification of the aortic and mitral valves from 4-d cardiac ct and tee. IEEE transactions on medical imaging **29**(9), 1636–1651 (2010)
8. 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)
9. 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)
10. Kachole, S.: Object Segmentation: From Neuromorphic Sensing to Neuromorphic Machine Learning. PhD thesis, Kingston University London (2024)
11. Kachole, S., Alkendi, Y., Naeini, F.B., Makris, D., Zweiri, Y.: Asynchronous events-based panoptic segmentation using graph mixer neural network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4083–4092 (2023)
12. Kachole, S., Huang, X., Naeini, F.B., Muthusamy, R., Makris, D., Zweiri, Y.: Bimodal SegNet: Fused instance segmentation using events and rgb frames. Pattern Recognition **149**, 110215 (2024)
13. Kachole, S., Hunter, G.J., Duran, O.: A computer vision approach to monitoring the activity and well-being of honeybees. In: Intelligent Environments (Workshops). pp. 152–161 (2020)
14. Kachole, S., Nayak, B., Brouner, J., Liu, Y., Guo, L., Makris, D.: Posture estimation from tactile signals using a masked forward diffusion model. Sensors **25**(16), 4926 (2025)
15. Kachole, S., Sajwani, H., Naeini, F.B., Makris, D., Zweiri, Y.: Asynchronous bioplausible neuron for spiking neural networks for event-based vision. In: European Conference on Computer Vision. pp. 399–415. Springer (2024)

16. Kachole, S., Thakur, S., Innani, S., Adap, S., You, S., Pitarch-Abaigar, C., Bakas, S.: Multi-frugal: Multimodal flexible redundancy-aware decomposed gated learning for cancer diagnosis and prognosis. arXiv preprint arXiv:2606.06867 (2026)
17. Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ayed, I.B.: Boundary loss for highly unbalanced segmentation. In: International conference on medical imaging with deep learning. pp. 285–296. PMLR (2019)
18. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896. Atlanta (2013)
19. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., et al.: Metrics reloaded: recommendations for image analysis validation. Nature methods **21**(2), 195–212 (2024)
20. Munafò, R., Saitta, S., Tondi, D., Ingallina, G., Denti, P., Maisano, F., Agricola, E., Votta, E.: Automatic 4d mitral valve segmentation from transesophageal echocardiography: a semi-supervised learning approach. Medical & Biological Engineering & Computing pp. 1–16 (2025)
21. Naeini, F.B., Kachole, S., Muthusamy, R., Makris, D., Zweiri, Y.: Event augmentation for contact force measurements. IEEE Access **10**, 123651–123660 (2022)
22. 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)
23. Pouch, A.M., Wang, H., Takabe, M., Jackson, B.M., Gorman III, J.H., Gorman, R.C., Yushkevich, P.A., Sehgal, C.M.: Fully automatic segmentation of the mitral leaflets in 3d transesophageal echocardiographic images using multi-atlas joint label fusion and deformable medial modeling. Medical image analysis **18**(1), 118–129 (2014)
24. Rehman, U., Hudson, H., Hao, C.Y., Ahn, Y., Kachole, S., Liu, J., Patel, S., Xu, M.J., Rouhani, M.J., O'Flynn, P., et al.: Lifetime prevalence of betel nut chewing in india and taiwan: Raising awareness of oral cancer risks and the urgent call for regulation. Cancers **18**(7), 1074 (2026)
25. Reinke, A., Tizabi, M.D., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Kavur, A.E., Rädsch, T., Sudre, C.H., Acion, L., Antonelli, M., et al.: Understanding metric-related pitfalls in image analysis validation. Nature methods **21**(2), 182–194 (2024)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
27. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PmLR (2019)
28. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
29. You, S., Pitarch-Abaigar, C., Kachole, S., Sonawane, S., Ha, J., Gada, A.S., Crandall, D., Shiradkar, R., Bakas, S.: PROFUSEme: PROstate cancer biochemical recurrence prediction via FUSEd multi-modal embeddings. arXiv preprint arXiv:2509.14051 (2025)
30. 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)