ValveMix: A Unified Semi-Supervised Framework for Multimodal Mitral Valve Anatomy Segmentation

Abdul Qayyum[1], Moona Mazher[2], and Steven A Niederer[1,3]

[1] National Heart and Lung Institute, Faculty of Medicine, Imperial College London, London, United Kingdom
[2] Hawkes Institute, Department of Computer Science, University College London, London, United Kingdom
[3]Division of Cardiovascular Medicine, Stanford Cardiovascular Institute, Standford University, CA, USA

Abstract. Accurate segmentation of mitral valve anatomy from multimodal medical imaging is essential for diagnosis, treatment planning, and intraoperative guidance in structural heart disease. However, robust segmentation across cardiac computed tomography (CT), three-dimensional transesophageal echocardiography (3D TEE), and surgical video remain challenging because of substantial differences in image characteristics and limited annotated data. We present ValveMix, a unified semi-supervised framework for multimodal mitral valve segmentation developed for the Mitral Valve Anatomy Analysis (MVAA) Challenge. The framework combines hierarchical convolutional feature extraction with UxLSTM bottleneck modules to capture both local anatomical details and long-range spatial dependencies. To leverage unlabeled CT and surgical video data, we incorporate a FixMatch-inspired semi-supervised strategy based on confidence-guided pseudo-labeling and consistency regularization. Five-fold cross-validation, model ensembling, and test-time augmentation are further employed to improve robustness. Compared with the official challenge baseline, the proposed framework improved the Dice similarity coefficient from 0.618 to 0.862 for cardiac CT, 0.745 to 0.850 for 3D TEE, and 0.644 to 0.743 for surgical video, while consistently reducing boundary errors. Semi-supervised learning further improved the Dice score for surgical video from 0.675 to 0.743. These results demonstrate the effectiveness of ValveMix for robust multimodal mitral valve segmentation and highlight the benefit of semi-supervised learning in scenarios with limited annotations.

Keywords: Mitral Valve Anatomy Analysis, MVAA Challenge, Cardiac CT, 3D Transesophageal Echocardiography, Surgical Video Segmentation, xLSTM-UNet, Semi-Supervised Learning, Test-Time Augmentation, Model Ensemble, Medical Image Analysis.

1 Introduction

Mitral valve disease is one of the leading causes of structural heart disease and affects millions of patients worldwide. Accurate assessment of mitral valve anatomy is essential for diagnosis, treatment planning, and image-guided interventions such as

transcatheter edge-to-edge repair (TEER), annuloplasty, and surgical valve repair. Modern clinical workflows rely on multiple complementary imaging modalities, each providing unique anatomical and functional information throughout different stages of patient management.

Cardiac computed tomography (CT) provides high-resolution three-dimensional anatomical information and is primarily used during pre-procedural planning for device sizing and anatomical assessment. Three-dimensional transesophageal echocardiography (3D TEE) serves as the principal intraoperative imaging modality, offering real-time visualization of valve morphology and function despite challenges arising from acoustic artifacts, noise, and limited image quality. During surgery, direct visualization through endoscopic or microscopic video provides complementary anatomical information that enables assessment of leaflet motion and procedural outcomes. Although deep learning has significantly advanced automatic medical image segmentation, most existing approaches are designed for a single imaging modality and do not address the heterogeneous characteristics of multimodal clinical data. The substantial differences in image appearance, spatial resolution, dimensionality, and acquisition protocols between CT, ultrasound, and surgical images present considerable challenges for developing robust segmentation algorithms. Consequently, there remains a need for methods capable of accurately segmenting mitral valve anatomy across diverse imaging environments while maintaining consistent anatomical representations.

The Mitral Valve Anatomy Analysis Using Multimodal Imaging Data (MVAA) Challenge [1] addresses this problem by providing a unified benchmark covering three clinically important imaging modalities spanning pre-operative, intraoperative, and surgical workflows. The challenge evaluates automated segmentation algorithms on cardiac CT, volumetric 3D TEE, and surgical video frames, encouraging the development of robust models that generalize across heterogeneous imaging domains. Participants may develop modality-specific models, while overall performance is determined by the combined results across the three tasks, reflecting complementary stages of the clinical workflow.

2 Proposed Method

This section describes the proposed xLSTM-UNet [2] framework for multimodal mitral valve segmentation across cardiac CT, 3D transesophageal echocardiography (3D TEE), and surgical video frames. We first introduce the challenge datasets and preprocessing pipeline, followed by the network architecture, the FixMatch-inspired semi-supervised learning strategy for exploiting unlabeled data, and the training and inference procedures, including five-fold cross-validation, model ensembling, and test-time augmentation.

2.1 Dataset

MVAA challenge dataset comprises multimodal mitral valve imaging collected from multiple clinical centers, including Guangdong Provincial People's Hospital, The First Affiliated Hospital of Jinan University, and Dongguan Eastern Central Hospital, with

ethics approval (No. KY-N-2022-048-01) [1]. The dataset includes three complementary imaging modalities spanning the clinical workflow: cardiac CT, 3D transesophageal echocardiography (3D TEE), and intraoperative surgical video frames. Cardiac CT data were acquired using scanners from Siemens, GE, Philips, and Canon, while 3D TEE images were obtained using Philips EPIQ ultrasound systems. The dataset consists of 1,067 CT volumes (27 labeled and 1,040 unlabeled), 105 labeled 3D TEE volumes, and 1,559 surgical video frames (180 labeled and 1,379 unlabeled), together with separate validation and test data released according to the challenge protocol. The labeled data were annotated through a two-stage expert review process involving multiple medical experts and senior radiologists, with quality control performed using inter-observer agreement analysis. All data were anonymized and provided in standardized formats, including NIfTI for CT and TEE and two-dimensional images for surgical video frames.

2.2 Preprocessing

All datasets were converted into a unified format while preserving their native spatial resolution, spacing, and orientation. Cardiac CT and 3D TEE volumes were automatically foreground-cropped and intensity-normalized, whereas surgical video frames were converted to RGB images and normalized channel-wise. Adaptive patch extraction was used according to the dimensionality and acquisition characteristics of each modality. Three-dimensional patches were used for CT and 3D TEE, while two-dimensional patches were extracted from the surgical frames. Standard spatial and intensity augmentations, including random rotation, scaling, mirroring, brightness, contrast, and gamma transformations, were applied during training. This unified preprocessing enabled the same segmentation framework to be applied across volumetric CT, ultrasound, and two-dimensional surgical images.

2.3 Model

2.3.1 xLSTM-UNet Architecture

We employ an enhanced xLSTM-UNet architecture that combines hierarchical convolutional feature extraction with UxLSTM bottleneck blocks for efficient long-range anatomical context modelling. The encoder consists of multiple convolutional stages that progressively extract multi-scale features while reducing spatial resolution through downsampling. At the bottleneck, conventional convolutional blocks are replaced by UxLSTM modules, which sequentially propagate information across the compressed feature representation to capture long-range dependencies between anatomically distant structures.

The decoder progressively reconstructs high-resolution segmentation maps by combining globally enriched bottleneck features with corresponding encoder representations through multi-scale skip connections. This feature fusion preserves fine anatomical boundaries while integrating global contextual information, enabling accurate delineation of thin mitral valve leaflets and annular structures. The same network design was employed for all challenge tasks. Three-dimensional models were trained for cardiac CT and volumetric 3D TEE segmentation, whereas a two-dimensional version was used

for surgical video frame segmentation to accommodate the different image dimensionalities.

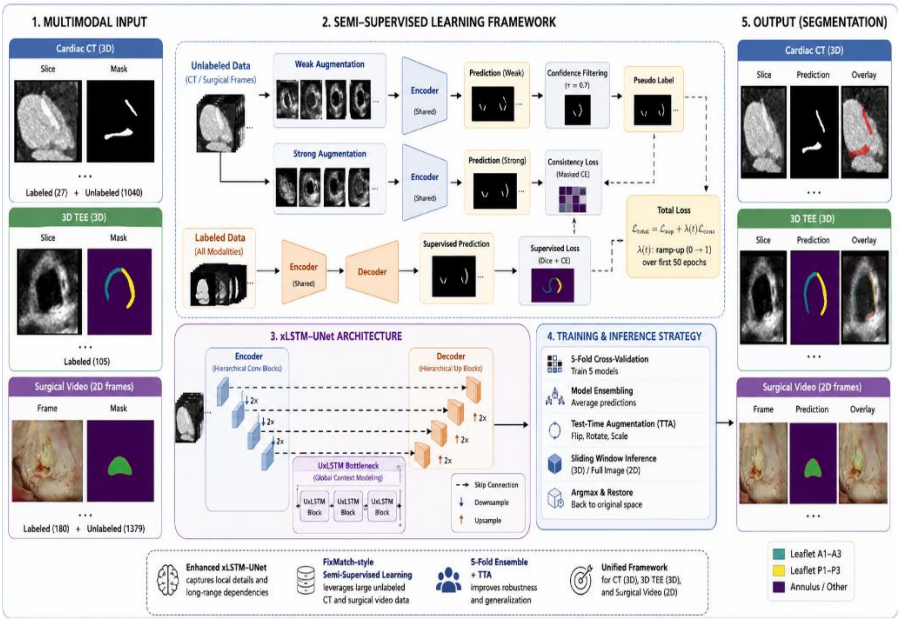

Figure 1. Overview of ValveMix, illustrating the unified semi-supervised learning pipeline, network architecture, training strategy, and representative segmentation results across cardiac CT, 3D TEE, and surgical video.

2.3.2 Hybrid Cardiac Loss and Cardiac-Specific Data Augmentation

To improve segmentation accuracy of the thin mitral valve anatomy, we introduce a hybrid segmentation loss that combines region-based, voxel-wise, hard-example, and boundary-aware supervision. The overall objective is defined as

$$L_{\text{Hybrid}} = L_{\text{Dice}} + L_{\text{CE}} + \lambda_1 L_{\text{TopK}} + \lambda_2 L_{\text{Boundary}}$$

where $\mathcal{L}_{Dice}$maximizes overlap between the predicted and reference segmentations, while $\mathcal{L}_{CE}$provides stable voxel-wise supervision. The TopK Cross-Entropy loss focuses optimization on the most difficult voxels by computing the loss over only the hardest predictions, improving learning under severe class imbalance. The boundary loss explicitly penalizes contour discrepancies and encourages accurate delineation of thin anatomical structures, including the mitral valve leaflets and annulus.

The weighting coefficients $\lambda_1$and $\lambda_2$control the contributions of the TopK and boundary losses, respectively. Throughout our experiments, we set $\lambda_1 = 0.5$and $\lambda_2 = 0.3$. Unlike conventional deep supervision, the TopK and boundary losses are applied only to the full-resolution prediction, while auxiliary decoder outputs are optimized using the

standard Dice and Cross-Entropy losses. This strategy improves boundary localization without destabilizing optimization at lower resolutions.

2.3.3 Semi-Supervised Learning

To exploit the large amount of unlabeled cardiac CT volumes and surgical video frames, we adopted a FixMatch-inspired consistency-based semi-supervised learning strategy [3]. During each training iteration, an unlabeled sample was processed using both weak and strong augmentations. The prediction generated from the weakly augmented image was converted into an online pseudo-label by selecting only pixels or voxels whose maximum class probability exceeded a confidence threshold of 0.7. These confident predictions were then used to supervise the corresponding strongly augmented sample through a masked consistency loss. Unlike conventional offline pseudo-labeling approaches that generate fixed labels before retraining, our framework continuously updates pseudo-labels throughout optimization, allowing supervision to improve as the segmentation model becomes progressively more accurate.

The overall optimization objective is defined as

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda(t)\mathcal{L}_{cons}$$

Where $\mathcal{L}_{sup}$denotes the supervised Dice and Cross-Entropy segmentation loss computed on labeled images,

$$\mathcal{L}_{sup} = \mathcal{L}_{Dice} + \mathcal{L}_{CE}$$

and $\mathcal{L}_{cons}$represents the confidence-masked consistency loss computed from unlabeled data. The consistency weight $\lambda(t)$is gradually increased during the first 50 training epochs using a sigmoid ramp-up schedule, reducing the influence of unreliable pseudo-labels during early optimization while fully exploiting unlabeled data in later training stages. Semi-supervised learning was applied to Task 1 (Cardiac CT) and Task 3 (Surgical Video), where substantial unlabeled datasets were available. Task 2 (3D TEE) was trained using supervised learning only because unlabeled training volumes were not provided.

2.3.4 Training and Inference

Each modality was trained independently using identical optimization settings. Five-fold cross-validation was performed for every task, producing five independently trained models. The best-performing checkpoint from each fold, selected according to the validation Dice score, was retained for inference. During testing, predictions from all five folds were combined through model ensembling [4-5]. Test-time augmentation (TTA) was additionally employed by averaging predictions obtained from multiple geometric transformations, resulting in more robust and stable segmentation masks. For volumetric CT and 3D TEE data, sliding-window inference was performed on the full image volume, whereas surgical video frames were processed individually using the two-dimensional network. Finally, voxel-wise or pixel-wise class probabilities were converted into segmentation masks using argmax classification and restored to the

original image space while preserving native resolution, spacing, and orientation. No additional task-specific post-processing was applied before challenge submission.

3 Results

Table 1 compares the official challenge baseline with the proposed ValveMix framework trained using supervised learning (ValveMix-SL) and semi-supervised learning (ValveMix-SSL). Both variants of the proposed framework substantially outperformed the baseline across all three modalities, demonstrating the effectiveness of the unified architecture for multimodal mitral valve segmentation. Compared with the baseline, ValveMix-SSL improved the Dice similarity coefficient from 0.618 to 0.862 for cardiac CT, from 0.745 to 0.850 for 3D TEE, and from 0.644 to 0.743 for surgical video, while consistently reducing the Hausdorff and average surface distances.

The ablation study further highlights the contribution of semi-supervised learning. Although only marginal improvements were observed for cardiac CT and 3D TEE, where sufficient labeled training data were available, semi-supervised learning provided a notable performance gain for surgical video, increasing the Dice score from 0.675 to 0.743 while reducing the Hausdorff distance from 106.65 to 97.68 and the average surface distance from 20.33 to 16.72. These results indicate that leveraging unlabeled data through confidence-guided pseudo-labeling and consistency regularization is particularly beneficial for challenging modalities with limited annotations and large appearance variability.

Table 1. Comparison of the challenge baseline and the proposed ValveMix framework with and without semi-supervised learning

Task	**Methods**	**DSC ↑**	**HD ↓**	**ASD ↓**
Cardiac CT	Baseline	0.618	9.621	0.959
	ValveMix-SL	0.860	4.449	0.268
	ValveMix-SSL	**0.862**	4.412	0.269
3D TEE	Baseline	0.745	18.767	1.655
	ValveMix-SL	0.847	10.744	0.605
	ValveMix-SSL	**0.850**	10.805	0.586
Surgical Video	Baseline	0.644	193.296	24.762
	ValveMix-SL	0.675	106.652	20.334
	ValveMix-SSL	**0.743**	**97.677**	**16.719**

The proposed method consistently outperformed the baseline across all three MVAA tasks. For Task 1 cardiac CT (30 cases), the Dice similarity coefficient increased from 0.618 to 0.862, while the Hausdorff distance decreased from 9.62 to 4.41 and the average surface distance from 0.96 to 0.27. This corresponds to a 54% reduction in FixMatchTEE (20 cases), the proposed model improved the Dice score from 0.745 to 0.850. The Hausdorff and average surface distances were reduced from 18.77 to 10.80 and from 1.66 to 0.59, respectively. For Task 3 surgical video segmentation (48 cases), the Dice score increased from 0.644 to 0.743, with the Hausdorff distance decreasing from 193.30 to 97.68 and the average surface distance from 24.76 to 16.72.

Qualitative comparisons support these quantitative improvements is shown in Figure 2. The proposed model generally produced more complete and anatomically coherent segmentations, with smoother boundaries and fewer false-positive regions than the baseline. The largest improvement in overlap accuracy was observed for cardiac CT, while substantial reductions in boundary errors were achieved across all three imaging modalities.

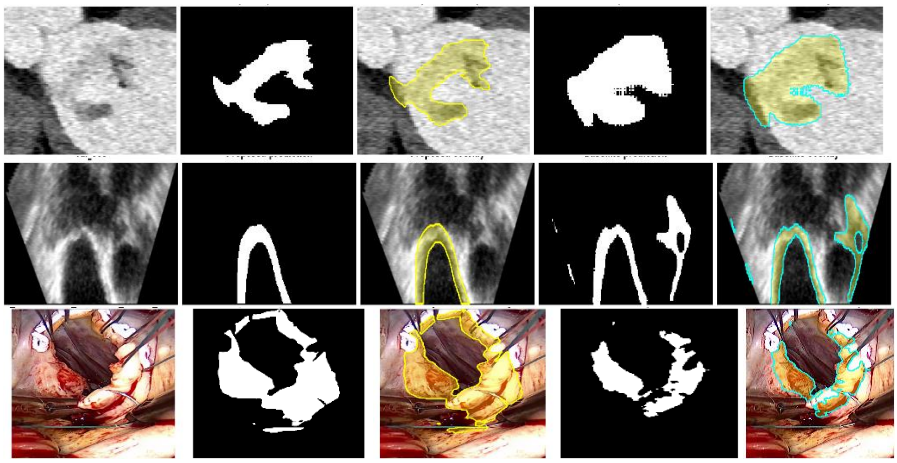

Figure 2. Qualitative comparison between the proposed ValveMix-SSL framework and the challenge baseline on representative cardiac CT, 3D TEE, and surgical video images. From left to right: input image, ValveMix-SSL prediction, ValveMix-SSL overlay (yellow), challenge baseline prediction, and challenge baseline overlay (cyan). The proposed method yields more accurate and anatomically coherent segmentations across all modalities.

4 Discussion

The proposed ValveMix framework demonstrated robust performance across three complementary imaging modalities, including cardiac CT, 3D transesophageal echocardiography (TEE), and surgical video. Despite substantial differences in image characteristics, dimensionality, and anatomical appearance, the unified framework consistently outperformed the official challenge baseline on all tasks. These results demonstrate that a shared architectural design can effectively learn modality-specific representations while maintaining strong generalization across heterogeneous imaging domains. Semi-supervised learning further improved segmentation performance by exploiting unlabeled cardiac CT and surgical video data through confidence-guided pseudo-labeling and consistency regularization. Although only modest improvements were observed for cardiac CT and 3D TEE, where sufficient labeled training data were available, substantial gains were achieved for the surgical video task. This suggests that semi-supervised learning is particularly beneficial for challenging scenarios characterized by limited annotations, large appearance variations, and complex intraoperative environments.

From a clinical perspective, accurate segmentation of mitral valve anatomy is essential for diagnosis, treatment planning, and image-guided structural heart interventions. The proposed framework produced anatomically coherent segmentations with smoother boundaries and fewer false-positive regions than the challenge baseline, indicating its potential to improve the reliability and reproducibility of automated mitral valve analysis. Nevertheless, this study has several limitations. The framework was evaluated only on the MVAA Challenge dataset, and semi-supervised learning was applied only to modalities with available unlabeled data. Future work will investigate larger multi-center datasets, self-supervised foundation model pretraining, and multimodal representation learning to further improve generalization and clinical applicability.

5 Conclusion

We presented ValveMix, a unified semi-supervised framework for multimodal mitral valve anatomy segmentation from cardiac CT, 3D TEE, and surgical video. By combining a shared segmentation architecture with confidence-guided pseudo-labeling, cross-validation ensembling, and test-time augmentation, the proposed framework consistently outperformed the official challenge baseline across all three tasks. Semi-supervised learning provided additional improvements, particularly for surgical video segmentation, demonstrating the effectiveness of leveraging unlabeled data in challenging clinical scenarios. These results highlight the potential of ValveMix as a robust and generalizable framework for automated mitral valve anatomy analysis and future computer-assisted structural heart interventions.

References

1. 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. MICCAI 2026. DOI: 10.5281/zenodo.19726755.
2. Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., ... & Hochreiter, S. (2024). xlstm: Extended long short-term memory. Advances in Neural Information Processing Systems, 37, 107547-107603.
3. Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. Advances in Neural Information Processing Systems 33, 596–608 (2020).
4. Mazher, M., Razzak, I., Qayyum, A., Tanveer, M., Beier, S., Khan, T., & Niederer, S. A. (2024). Self-supervised spatial–temporal transformer fusion based federated framework for 4D cardiovascular image segmentation. Information Fusion, 106, 102256.
5. Qayyum, Abdul, Moona Mazher, Devran Ugurlu, Jose Alonso Solis Lemus, Cristobal Rodero, and Steven A. Niederer. "Foundation model for whole-heart segmentation: Leveraging student-teacher learning in multi-modal medical imaging." arXiv preprint arXiv:2503.19005 (2025).