When Does Unlabelled Data Help? A Shift-Aware Semi-supervised Study of Multimodal Mitral Valve Segmentation

Sidratul Montaha[1], Nasif Hossain[1], Michelle Noga[2], Kumaradevan Punithakumar[2], and Nilanjan Ray[1]

[1] Department of Computing Science, University of Alberta, Edmonton, AB, Canada
[2] Department of Radiology and Diagnostic Imaging, University of Alberta, Edmonton, AB, Canada
{montaha,nhossain,mnoga,punithak,nray1}@ualberta.ca

Abstract. Semi-supervised learning is commonly used in medical image segmentation with limited annotations, yet its benefit may depend on the underlying data distribution. We examine this question within a multimodal mitral valve segmentation study using the MVAA challenge, where CT and surgical video include unlabelled data and ultrasound provides a fully supervised reference. We analyse three factors: unlabelled-data volume, the labelled–unlabelled data gap, and acquisition-level variation within the semi-supervised tasks. For CT, where the measured data gap and acquisition-level variation are small, we train a 3D SegResNet ensemble, generate pseudo-labels for unlabelled volumes with flip test-time augmentation (flip-TTA), retain only masks passing confidence–uncertainty and anatomical-size gates, and retrain the ensemble together with a warm-started Bidirectional Copy-Paste (BCP) EMA-teacher model. For ultrasound, which has no unlabelled data, we use a fully supervised heterogeneous nnU-Net ensemble. For surgical video, where the labelled–unlabelled gap and acquisition-level variation are larger, we use a frozen DINOv2 backbone with mean-teacher pseudo-labelling and appearance randomisation, accepting pseudo-labels only when confidence and mask coverage are plausible.
Reliable pseudo-labelling benefits both semi-supervised tasks (about 3% Dice improvement on CT and 5% on video), whereas foundation features and appearance robustness yield substantial gains only on high-variation surgical video and are marginal on low-variation CT. Overall, our findings suggest that the value of unlabelled data in multimodal segmentation depends on where it adds reliable, task-relevant variation.

Keywords: Mitral valve segmentation · Semi-supervised learning · Distribution shift · Foundation models · Multimodal segmentation.

1 Introduction

The mitral valve is the most commonly repaired heart valve, and mitral regurgitation is among the most prevalent valvular heart diseases, with prevalence

rising with age [5,18]. Successful repair relies on accurate delineation of the valve across the surgical pathway: preoperative cardiac CT for anatomical assessment, intraoperative three-dimensional transesophageal echocardiography (TEE) for real-time guidance, and surgical video for operative-field analysis. Because manual delineation in these modalities is slow, operator-dependent, and hard to scale, automatic mitral valve segmentation across CT, ultrasound, and video is a clinically relevant goal.

Prior mitral valve segmentation work has focused mainly on single-modality echocardiography, including 3D TEE residual U-Nets [2], multi-decoder residual networks [12], two-stage nnU-Net pipelines [4], and 2D leaflet tracking [21]. ImageCAS is a large CT benchmark for cardiovascular segmentation [23]. In contrast, valve segmentation in surgical video remains little studied, with existing work addressing related endoscopic scenes rather than direct valve segmentation [8]. Because expert annotations are limited, semi-supervised learning is a natural option, using either consistency training or pseudo-labelling [11,9,20,22]. Recent work in 4D TEE shows its promise [13], while foundation models and appearance-randomisation methods may improve robustness to acquisition differences [3,15,16]. However, unlabelled data and stronger backbones are not equally useful in every setting. When recordings differ in illumination, contrast, texture, or acquisition conditions, performance may be limited more by distribution shift than by label scarcity.

We study mitral valve segmentation across CT, ultrasound, and surgical video to examine when semi-supervision is useful. Our contributions are: (1) a unified multimodal analysis of how unlabelled-data volume, labelled–unlabelled gap, and acquisition-level shift affect performance; (2) modality-specific systems for CT, ultrasound, and video; and (3) a comprehensive evaluation showing that pseudo-labelling benefits CT and video, while foundation features and appearance robustness mainly help high-shift surgical video.

2 Methods

We study when unlabelled data improves mitral valve segmentation, starting from a supervised baseline with controlled additions. In parallel, we quantify the labelled–unlabelled data gap and acquisition-level shift of each task. Let $\mathcal{D}^m = \mathcal{D}_L^m \cup \mathcal{D}_U^m$ denote the data for modality $m \in \mathrm{CT}, \mathrm{US}, \mathrm{VID}$, where $\mathcal{D}_L^m = (x_i, y_i)$ is the labelled set and $\mathcal{D}_U^m = x_j$ is the unlabelled set when available. Here, x is a CT volume, an ultrasound volume, or a video frame, and y is the binary mitral-valve mask. CT and video include unlabelled data, whereas ultrasound does not.

2.1 Baseline Segmentation Models

Cardiac CT. For cardiac CT, we use a 3D SegResNet [14], matching the volumetric structure of the valve and its continuity across slices. The network processes 96^3 sub-volumes and uses an encoder with $\{1, 2, 2, 4\}$ residual blocks at

feature widths $\{32, 64, 128, 256\}$, followed by a symmetric decoder. CT intensities are clipped to $[-200, 800]$ HU, rescaled to $[0, 1]$, and normalised over non-zero voxels. The model is trained with a combined soft Dice and cross-entropy loss. At inference, we use overlapping sliding-window prediction with Gaussian weighting and retain the largest connected component.

3D Ultrasound. For 3D transesophageal echocardiography, we use nnU-Net [7] as the supervised reference because it automatically adapts preprocessing, patch size, voxel spacing, and network configuration to the dataset. We train the 3D full-resolution model using five patient-disjoint folds. To strengthen the final ultrasound system, we also train a residual-encoder variant, `ResEncUNet-L`, under the same folds and average the predicted probabilities across both configurations. Since ultrasound has no unlabelled data, this supervised ensemble defines the final ultrasound model and serves as the modality-specific comparison point.

Surgical video. For surgical video, we explicitly compare a conventional convolutional encoder with a frozen self-supervised foundation backbone under the same decoder and training protocol. The convolutional reference uses a U-Net decoder with an EfficientNetV2-M encoder. The foundation variant replaces the encoder with a frozen DINOv2 ViT-L/14 backbone [15], so only the decoder is trained on the labelled frames. Frames are processed at 644×1148, matching the DINOv2 patch size, and both models are trained with Dice and boundary-weighted cross-entropy losses. The convolutional model provides the standard video reference, while the DINOv2 model is used as the stronger supervised baseline for the semi-supervised and appearance-robustness stages.

2.2 Measuring Distribution Shift

We quantify shift at the acquisition level, since frames from the same video and slices from the same study are highly correlated. For each modality, we hold out one acquisition group at a time, train the baseline on the remaining groups, and measure the Dice score s_g on the held-out group g. The across-group spread is

$$\sigma^m = \sqrt{\frac{1}{G^m}\sum_{g=1}^{G^m}(s_g - \bar{s})^2}, \qquad \bar{s} = \frac{1}{G^m}\sum_{g=1}^{G^m} s_g. \tag{1}$$

A larger σ^m indicates less stable generalisation across acquisitions. We also report worst-group Dice, following worst-group generalisation [17].

We additionally compute label-free image-level separability using Fréchet distance [6] and a normalised $\mathcal{A}$-distance classifier [1,10]. Samples are embedded with a frozen DINOv2 encoder, and a linear classifier is trained to distinguish two data pools. If BAcc is the balanced accuracy, we define

$$d_{\mathcal{A}} = 2\max(\mathrm{BAcc}, 1 - \mathrm{BAcc}) - 1, \tag{2}$$

where 0 indicates indistinguishable pools and 1 indicates perfect separability. For CT and video, the same measure is used to estimate the labelled–unlabelled data gap by classifying labelled versus unlabelled samples.

Finally, we define task-Jacobian sensitivity as a task-aware, label-free shift measure. For an input x and segmenter f_θ, we compute the foreground response and its input-gradient magnitude as

$$S(x) = \sum \mathrm{ReLU}(f_\theta(x)), \qquad J(x) = \|\nabla_x S(x)\|_1. \tag{3}$$

The sensitivity score is averaged within each acquisition group. Unlike generic image-level distances, this measure reflects input variation in directions that affect the predicted segmentation.

Overall, label-free measures indicate image-level mismatch, however they do not directly show whether the mismatch affects segmentation. When the two disagree, we prioritise group-disjoint σ^m, as it directly quantifies performance variation across held-out acquisition groups. Thus, label-free metrics are used as diagnostic indicators, while σ^m guides robustness interpretation when labelled validation data are available.

2.3 Proposed Approaches

The three modality-specific systems reflect the different data settings observed in the MVAA tasks. For CT, where labelled data is limited but acquisition-level variation is comparatively low, we use conservative pseudo-labelling with reliability and plausibility checks. For surgical video, where acquisition-level variation is higher, we combine foundation features, selective pseudo-labelling, and appearance randomisation. This provides a consistent framework for studying how data characteristics relate to the benefit of semi-supervised and robustness-oriented methods.

CT Segmentation with Reliability-Gated Pseudo-Labelling. In the low-label semi-supervised setting, a single model can produce confident however incorrect masks, and using these masks for retraining may reinforce its own errors. We therefore use a reliability- and plausibility-gated pseudo-labelling strategy: an unlabelled CT volume is added to training only when the predicted mask is both internally reliable and anatomically plausible. We first train a five-fold supervised SegResNet ensemble on the labelled CT data using patient-disjoint folds. For each unlabelled volume x_j, the ensemble prediction is averaged across folds and flip-based test-time augmentation,

$$\bar{p}(x_j) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{2} \Big(f_{\theta_k}(x_j) + \mathcal{F}[f_{\theta_k}(\mathcal{F}[x_j])] \Big), \tag{4}$$

where $K = 5$ and $\mathcal{F}$ denotes the spatial flip. The pseudo-label is obtained as $\hat{y}_j = \mathbb{1}[\bar{p}(x_j) > 0.5]$, with foreground region $\Omega_j = \{v : \hat{y}_j(v) = 1\}$.

To estimate pseudo-label reliability, we compute the mean foreground confidence and entropy:

$$\mathrm{conf}(x_j) = \frac{1}{|\Omega_j|} \sum_{v \in \Omega_j} \bar{p}_v, \qquad \mathrm{unc}(x_j) = \frac{1}{|\Omega_j|} \sum_{v \in \Omega_j} H(\bar{p}_v), \tag{5}$$

where $H(p) = -p\log p - (1-p)\log(1-p)$. These terms are combined as $r(x_j) = \mathrm{conf}(x_j)\big(1 - \mathrm{unc}(x_j)\big)$, so a pseudo-label must be both confident and low-uncertainty to receive a high score.

Reliability alone does not guarantee anatomical validity. We therefore add a size-based plausibility gate using the foreground-volume range observed in the labelled CT set. A pseudo-label is accepted only if $r(x_j) \geq \tau$ and $V_{\min} \leq |\Omega_j| \leq V_{\max}$, where $\tau = 0.6$ and $[V_{\min}, V_{\max}]$ is computed from the labelled masks. This rejects empty masks, severe leakage into neighbouring structures, and implausibly fragmented predictions.

The accepted pseudo-labelled set $\mathcal{P}$ is then combined with the labelled data for retraining, while validation in each fold uses only real labelled cases to avoid pseudo-label leakage. In parallel, we train a bidirectional copy-paste teacher that mixes labelled and unlabelled sub-volumes with a random cuboid mask M as $x_{\mathrm{mix}} = M \odot x_l + (1-M) \odot x_u$ and $y_{\mathrm{mix}} = M \odot y_l + (1-M) \odot \hat{y}_u$. The teacher is maintained as an exponential moving average of the student, $\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s$ with $\alpha = 0.99$, and is warm-started from the supervised model to avoid unstable early pseudo-labels. The final CT system combines the retrained pseudo-label ensemble with the BCP teacher model.

Ultrasound Segmentation with a Supervised Ensemble. The ultrasound task is multi-class, with the mitral valve annotated as anterior leaflet, posterior leaflet, and background. Because the two leaflets are adjacent and can be difficult to separate, we use a supervised ensemble to improve robustness over a single model. We train two nnU-Net-based configurations: the standard nnU-Net [7] and a deeper residual-encoder variant, ResEncUNet-L. Each configuration is trained with five patient-disjoint folds, giving ten models in total. The deployed Task 2 submission uses this ten-model ensemble: at inference, we average the per-class softmax probabilities across all models and assign each voxel to the class with the highest averaged probability. Since no unlabelled ultrasound data is available, this supervised ten-model ensemble serves as both the ultrasound baseline and the final Task 2 system.

Video Segmentation with Foundation Features and Appearance-Robust Training. Surgical video is found as the highest-shift and the most challenging modality, with illumination, contrast, tissue appearance, and valve visibility varying between operations. Here we combine three components: a frozen foundation backbone, plausibility-gated pseudo-labelling for the unlabelled frames, and appearance randomisation during training.

Foundation backbone. We use a self-supervised DINOv2 ViT-L/14 encoder [15] with a trained decoder. For the baseline and all controlled experiments, the backbone is kept frozen, since fine-tuning a large encoder on only 180 labelled frames overfits, whereas the fixed self-supervised features are already robust to the appearance variation. The decoder is trained at three input resolutions (504×896, 588×1036, 672×1190) to capture the valve at multiple scales; the final ensemble additionally includes variants in which the backbone is fine-tuned at a reduced learning rate, for architectural diversity.

Plausibility-gated pseudo-labelling. We generate pseudo-labels with a teacher ensemble trained on the labelled frames, a ConvNeXt-based UNet++ and the DINOv2 segmenter—whose predictions on each of the unlabelled frames are averaged. To avoid accepting unreliable masks, we measure confidence near the predicted valve rather than over the full image. The mask is dilated with a 15×15 kernel to define a local band $\mathcal{B}$, and confidence is computed as $\text{conf} = \frac{1}{|\mathcal{B}|}\sum_{v\in\mathcal{B}} 2\big|p_v - 0.5\big|$. We also apply a size constraint using the predicted foreground fraction $c = \frac{1}{HW}\sum_v \hat{y}_v$. A frame is accepted only if $\text{conf} \geq 0.15$ and $0.005 \leq c \leq 0.6$, which rejects near-empty masks and severe over-segmentations that commonly occur when the valve is poorly visible or occluded. This retains 1069 of the 1379 frames, which are added to the labelled set with loss weight 0.5 for student retraining.

Appearance randomisation. To reduce dependence on recording-specific texture and contrast, we apply GIN-style appearance randomisation during training. A random grouped-convolutional transform g_ϕ modifies local appearance and is blended with the original image as $\tilde{x} = \alpha\, g_\phi(x) + (1-\alpha)\, x$, with $\alpha \sim \mathcal{U}(0.3, 1.0)$. The transform is applied with probability 0.8 and resampled for each image. It changes appearance while preserving spatial structure, so the original segmentation mask remains valid. This encourages the model to segment the valve from structure rather than recording-specific visual cues.

The final system is a five-model ensemble. It includes two EfficientNet-based convolutional segmenters [19] with UNet++ and DeepLabV3+ decoders at 384×640, and three DINOv2 ViT-L/14 models at 504×896, 588×1036, and 672×1190, with the two higher-resolution DINOv2 models trained using appearance randomisation. Predictions are averaged with horizontal-flip test-time augmentation, followed by connected-component cleanup and hole filling.

3 Result Analysis

3.1 Experimental Setup

Data and Evaluation Protocol. We evaluate three MVAA tasks: cardiac CT (27 labelled, 1040 unlabelled volumes), 3D ultrasound (105 labelled studies), and surgical video (180 labelled frames from six recordings, 1379 unlabelled). Internal experiments use group-disjoint cross-validation to prevent acquisition-level leakage. We report Dice, σ, HD95, and ASD internally, and Dice, HD, and ASD on the hidden test set.

Implementation. All models are trained with AdamW, a cosine learning-rate schedule, and a Dice–cross-entropy objective. The CT SegResNet is trained for 150 epochs, with EMA decay $\alpha = 0.99$ for the copy-paste teacher. The ultrasound nnU-Net models are trained for 200 epochs at 3D full resolution. The video models use a DINOv2 ViT-L/14 backbone with a trained decoder, frozen for the baseline and analysis models, and fine-tuned at a reduced learning rate (0.1× the decoder rate) for some ensemble members—trained for 60–80 epochs with cosine warmup at input resolutions from 504×896 to 672×1190. Appearance

Table 1: Baseline performance and data characteristics per modality. σ and "Worst" report performance variation across held-out groups; "Sens." is task-Jacobian input sensitivity; and "Gap" is the labelled–unlabelled $\mathcal{A}$-distance.

Modality	Backbone	Dice	σ	Worst	Sens.	Gap
CT	SegResNet (3D)	0.816	0.070	0.62	0.3	0.22
Ultrasound	nnU-Net (3D)	0.837	–	–	–	–
Video	EfficientNetV2-M	0.699	0.153	0.38	38.7	0.90

randomisation is applied with probability 0.8. All models are implemented in PyTorch with MONAI and trained on RTX 4090 GPUs.

3.2 Baseline Performance and Distribution Shift

Baseline Performance. Table 1 reports baseline accuracy and shift per modality. CT and ultrasound are accurate and stable across held-out studies: the CT SegResNet reaches 0.816 Dice with a low spread (σ=0.07), and ultrasound reaches 0.837 with nnU-Net. Surgical video is far less stable—the convolutional (EfficientNetV2-M) baseline reaches only 0.699 Dice and varies strongly across recordings (σ=0.15, range 0.38–0.84), so its accuracy depends heavily on the held-out recording.

Distribution shift. The task-aware measures identify surgical video as substantially higher shift than CT: it has a larger performance spread ($\sigma = 0.15$ vs. 0.07) and much higher input sensitivity (38.7 vs. 0.3). In contrast, the image-level $\mathcal{A}$-distance gives the same value for both modalities (0.78), suggesting that it captures general appearance differences rather than segmentation difficulty. The labelled–unlabelled gap shows a similar pattern, being larger for video (0.90) than for CT (0.22).

3.3 Shift-Dependent Effect of Each Strategy

We next evaluate whether the measured data characteristics explain the effect of each intervention. The results indicate early saturation with increasing unlabelled-data volume. Using unfiltered pseudo-labels, CT achieves most of its internal improvement with 25% of the pool (0.816 $\rightarrow$ 0.837), whereas video saturates at approximately 50% (0.699 $\rightarrow$ 0.723). Beyond these levels, additional data changes Dice by less than 1%. These unfiltered internal gains did not fully transfer to the hidden test; our final systems (section 2.3) instead apply reliability- and plausibility-gated pseudo-labelling to all unlabelled data, which yields the results in Table 2.

Table 2 shows the interventions affect CT and video in distinct ways. Reliability- and plausibility-gated pseudo-labelling improves both modalities, with gains of +0.029 Dice in CT and +0.051 in video, consistent with the limited number of annotations in both tasks. In contrast, foundation features and appearance randomisation mainly benefit surgical video. Replacing the convolutional encoder

Table 2: Cumulative ablation under internal group-disjoint validation and external hidden-test evaluation. Each row adds one strategy to the previous configuration. DSC, σ, HD95, and ASD are reported with changes from the preceding row; external values are from the final second-phase CodaBench submission.

	Internal (group-disjoint)					External (hidden test)		
Configuration	DSC	Δ	σ	HD95	ASD	DSC	HD95	ASD
Surgical video — high shift ($\sigma_{\text{base}} = 0.15$)								
CNN baseline	0.699	—	0.153	86.9	26.8	–	–	–
→ DINOv2	0.751	↑.052	0.050	52.1	17.6	–	–	–
→ +pseudo	**0.802**	↑.051	**0.028**	52.6	**16.0**	–	–	–
→ +GIN (final)	0.771	↓.031	0.051	52.4	16.5	**0.817**	**236.28**	**69.05**
Cardiac CT — low shift ($\sigma_{\text{base}} = 0.07$)								
SegResNet baseline	0.816	—	0.070	2.44	0.64	–	–	–
→ +pseudo (final)	**0.845**	↑.029	0.061	**2.10**	**0.57**	**0.831**	**5.74**	**0.53**
→ +int.-norm	0.842	↓.003	0.060	2.11	0.57	–	–	–
Ultrasound/TEE — supervised reference task								
Final supervised ensemble	–	–	–	–	–	**0.856**	**9.91**	**0.51**

with DINOv2 improves video Dice by 0.052 and reduces the spread from $\sigma = 0.153$ to $\sigma = 0.050$, whereas comparable robustness-oriented additions provide little benefit in stable CT. These results match the earlier shift measurements: methods that reduce input sensitivity produce larger gains in high-shift video, while pseudo-labelling helps both tasks by addressing annotation scarcity.

Confidence–uncertainty filtering achieved the highest partial internal five-fold mean CT Dice of 0.8896, compared with 0.8566 for single-forward pseudo-labelling, but did not improve external CodaBench performance. BCP EMA-teacher provided only modest local refinement, improving one pseudo-retrained fold from 0.8570 to 0.8614 and STU-Net fold 0 from 0.8566 to 0.8666. The pseudo-labeling gain was +0.029 Dice for CT (approximate 95% confidence interval [CI], −0.067 to 0.125) and +0.051 for surgical video (approximate 95% CI, −0.003 to 0.105).

4 Conclusion

We studied mitral valve segmentation across CT, 3D ultrasound, and surgical video to examine when unlabelled data and stronger models are useful. The results show that semi-supervised learning gives early gains, but its value depends on the labelled–unlabelled data gap and the acquisition-level shift of the task. The proposed task-Jacobian sensitivity helps identify performance-relevant shift where generic image distances fail. Semi-supervision and appearance robustness are most beneficial for high-shift surgical video, while stable CT gains little after early saturation. These findings support a practical rule for multimodal medical segmentation: measure the data gap and task shift before adding unlabelled data or model capacity.

References

1. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning **79**(1), 151–175 (2010)
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. Cekmeceli, K., Himmetoglu, M., Tombak, G.I., Susmelj, A., Erdil, E., Konukoglu, E.: Do vision foundation models enhance domain generalization in medical image segmentation? In: European Conference on Computer Vision. pp. 185–200. Springer (2024)
4. Chen, J., Li, H., He, G., Yao, F., Lai, L., Yao, J., Xie, L.: Automatic 3d mitral valve leaflet segmentation and validation of quantitative measurement (2023)
5. Coffey, S., Roberts-Thomson, R., Brown, A., Carapetis, J., Chen, M., Enriquez-Sarano, M., Zühlke, L., Prendergast, B.D.: Global epidemiology of valvular heart disease. Nature Reviews Cardiology **18**(12), 853–864 (2021)
6. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
7. 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)
8. Ivantsits, M., Tautz, L., Sündermann, S., Wamala, I., Kempfert, J., Kuehne, T., Falk, V., Hennemuth, A.: Dl-based segmentation of endoscopic scenes for mitral valve repair. In: Current Directions in Biomedical Engineering. vol. 6, p. 20200017. De Gruyter (2020)
9. Jiao, R., Zhang, Y., Ding, L., Xue, B., Zhang, J., Cai, R., Jin, C.: Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. Computers in Biology and Medicine **169**, 107840 (2024)
10. Lopez-Paz, D., Oquab, M.: Revisiting classifier two-sample tests. arXiv preprint arXiv:1610.06545 (2016)
11. Montaha, S., Rahman, R., Ghosh, T., Sheikhi, F., Maleki, F.: Federated pseudo-labeling: A data-centric, privacy-preserving framework for medical image segmentation. IEEE journal of biomedical and health informatics **29**(12), 8823–8830 (2025)
12. Munafo, R., Saitta, S., Ingallina, G., Denti, P., Maisano, F., Agricola, E., Redaelli, A., Votta, E.: A deep learning-based fully automated pipeline for regurgitant mitral valve anatomy analysis from 3d echocardiography. IEEE Access **12**, 5295–5308 (2024)
13. 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 (2025)
14. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: MICCAI Brainlesion Workshop. pp. 311–320. Springer (2018)
15. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
16. Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., Rueckert, D.: Causality-inspired single-source domain generalization for medical image segmentation. IEEE Transactions on Medical Imaging **42**(4), 1095–1106 (2022)

17. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
18. Singh, J.P., Evans, J.C., Levy, D., Larson, M.G., Freed, L.A., Fuller, D.L., Lehman, B., Benjamin, E.J.: Prevalence and clinical determinants of mitral, tricuspid, and aortic regurgitation (the framingham heart study). The American journal of cardiology **83**(6), 897–902 (1999)
19. Tan, M., Le, Q.V.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning (ICML). pp. 6105–6114 (2019)
20. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results (2017)
21. Wifstad, S.V., Kildahl, H.A., Grenne, B., Holte, E., Hauge, S.W., Sæbø, S., Mekonnen, D., Nega, B., Haaverstad, R., Estensen, M.E., et al.: Mitral valve segmentation and tracking from transthoracic echocardiography using deep learning (2024)
22. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: MICCAI (2019)
23. Zeng, A., Wu, C., Lin, G., Xie, W., Hong, J., Huang, M., Zhuang, J., Bi, S., Pan, D., Ullah, N., et al.: Imagecas: A large-scale dataset and benchmark for coronary artery segmentation based on computed tomography angiography images. Computerized Medical Imaging and Graphics **109**, 102287 (2023)