Vendor-Agnostic 3D Mitral Valve Segmentation Using Semi-Supervised Learning

Maïlys Hau
IDI, *NTNU*
Trondheim, Norway
0000-0002-3449-7268

Riccardo Munafò
DEIB, *Politecnico di Milano*
Milan, Italy
0000-0002-5624-2776

Simone Saitta
Department of Biomedical Engineering and Physics
Amsterdam UMC, Informatics Institute, University of Amsterdam
Amsterdam, The Netherlands
DEIB, Politecnico di Milano
Milan, Italy
0000-0002-3974-5945

Federico Veronesi
GE Vingmed Ultrasound
Horten, Norway
0000-0003-3774-206X

Giacomo Ingallina
Unit of Cardiovascular, *IRCCS, San Raffaele Hospital*
Milan, Italy
0000-0002-4102-6405

Bjørnar Grenne
ISB, *NTNU*
Clinic of Cardiology, St Olavs University Hospital
Trondheim, Norway
0000000229846865

Frank Lindseth
IDI, *NTNU*
Trondheim, Norway
0000-0002-4979-9218

Emiliano Votta
DEIB, *Politecnico di Milano*
Milan, Italy
0000-0001-7115-0151

Gabriel Kiss
IDI, *NTNU*
Trondheim, Norway
0000-0001-5024-1548

2025 IEEE Symposium on Computational Intelligence in Health and Medicine (CIHM) | 979-8-3315-0833-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/CIHM64979.2025.10969485

***Abstract*—Mitral valve assessment for disease diagnosis and treatment is commonly guided by ultrasound imaging, with 3D transesophageal echocardiography being the de facto standard modality. The complex 3D structure of the mitral valve poses a challenge for its time-efficient and accurate quantification. Deep neural networks have been proposed for automatic mitral valve segmentation, but variations in imaging devices and protocols across ultrasound vendors cause significant domain differences in echocardiography datasets, leading to suboptimal dataset-dependent segmentation techniques.**

In this work, we apply a semi-supervised learning method to develop a vendor-agnostic, automatic segmentation tool for 3D transesophageal echocardiography. We learn from one manually annotated ultrasound domain, to generate reliable pseudo-labels using an ensemble of three models, for another unseen ultrasound domain. Our teacher-student framework is validated on images from two of the biggest cardiac ultrasound manufacturers.

In our experiments, the student model outperforms each teacher, with a Dice score of 82.20 ± 4.62% on known data and 71.32 ± 6.95% on unseen data, and an Average Surface Distance of 0.37 ± 0.07 mm and 0.82 ± 0.15 mm for the mitral valve on known and unknown domain respectively. These results enable efficient cross-domain analysis by reducing the need for manual annotation and ensuring consistent mitral valve analysis across different vendors.

This work was supported by the Centre for Innovative Ultrasound Solutions (CIUS) and the Department of Computer Science, Norwegian University of Science and Technology (NTNU) and the European Union's Horizon 2020 research and innovation program under the project ARTERY, grant No. 101017140, Politecnico di Milano.

(Maïlys Hau and Riccardo Munafò are co-first authors) (Corresponding authors: Maïlys Hau; Riccardo Munafò)

***Index Terms*—3D segmentation, ultrasound, semi-supervised, mitral valve**

I. Introduction

Mitral valve (MV) diseases affect approximately 2% of the general population [1]. Proper management requires diagnosis, screening, careful planning, and follow-up, which typically rely on Ultrasound (US) imaging for MV assessment. In particular, transthoracic echocardiogram (TTE) can be used for a first evaluation, but for a thorough diagnosis transesophageal echocardiography (TEE) is required [2]. TEE is also commonly used to monitor the MV repair intra-operatively, offering real-time 2D and 3D visualization of the MV, allowing for a detailed evaluation of the valve morphology. However, if manually performed, quantitative assessment of the MV requires pixel/voxel-level annotations, a labour-intensive and time-consuming process that often exceeds the time constraint of standard examinations, even for experienced sonographers [3], and it is not compatible with real-time use.

Machine learning (ML) algorithms, particularly those based on convolutional neural networks (CNNs), enable automatic MV segmentation, providing reliable, repeatable, and time-efficient analysis. The performance of these methods heavily relies on the amount and quality of the labeled data used for their training. Moreover, the ability of these methods to generalize to unseen data is often evaluated on test sets with similar distributions as the training sets. In echocardiography, this translates to CNNs trained and evaluated on single-centre

datasets where data are acquired using a single ultrasound scanner. In fact, variations in imaging devices and protocols between US manufacturers result in significant domain differences in echocardiography datasets, e.g. in terms of speckle pattern, intensity distribution, or contrast. These differences can degrade the performance of segmentation networks on out-of-distribution echocardiography samples, and hence prevent their translation into real-world clinical tools.

II. Related Work

CNNs have shown promising performance for supervised MV segmentation from cardiac US. In [4] the UNet architecture was applied for MV segmentation from 2D images. Later, several UNet variants have been proposed, such as 3D Residual UNet [5], Multi-Decoder Residual UNet [6], SegResNet [7] and nnUNet [8], [9], for MV segmentation in its closed state [6], [8], open state [5], [7] or both [9] from 3D TEE images. More recently, a semi-supervised training strategy with pseudo-labeling [10] was proposed to achieve automatic 4D segmentation of the MV: by leveraging only the MV ground truth (GT) segmentations at end-systole and end-diastole frames — respectively closed and open MV configurations.

Despite these considerable efforts, all of these works are limited to volumetric data acquired using single-vendor US scanners. As a result, the performance of the proposed segmentation networks on images from other US manufacturers remains to be evaluated. This challenge could potentially be mitigated through domain adaptation models, such as generative advesarial Networks (GANs) [11] or diffusion models [12], which enable images from different domains to appear visually similar while preserving semantic content. Although domain adaptation techniques have been explored in medical imaging [13], their application to cardiac US is limited and focuses only on 2D imaging [14] [15]. Despite their potential effectiveness, these methods are memory-intensive, difficult to train, and often less effective on low-contrast images, making them unsuitable for domain adaptation in 3D TEE segmentation.

This work introduces a self-training strategy for cross-domain MV segmentation from 3D TEE. Inspired by [10], we apply a semi-supervised learning (SSL) strategy based on a Teacher-Student framework, that is now extended to handle data coming from two different US vendors. We leverage GT MV segmentations of a single echocardiography domain to train a teacher network, which was subsequently used to create pseudo labels for another domain. Finally, a student network was trained on the combined domains using manual and pseudo labels. This approach offers a viable and effective alternative to high-demand domain adaptation methods, requiring only a marginal increase in cost compared to a traditional fully-supervised learning approach.

III. Methods

A. Dataset

Two hundred and forty (240) MV-centred 4D TEE data, i.e., 3D volumes acquired over the whole cardiac cycle, from 151 subjects were collected from St. Olavs Hospital in Trondheim, Norway, and IRCCS San Raffaele Hospital in Milan, Italy. Both studies were approved by local ethical committees and all patients provided informed consent. Exactly half of the total dataset was acquired with a Vivid E95 system with a 6VT probe (General Electric HealthCare Vingmed Ultrasound, Horten, Norway) on 91 subjects, including healthy individuals and patients with MV regurgitation or MV stenosis. The other half was acquired with a Philips EPIQ CVx system with an X8-2T probe (Philips, Andover, MA, US) on 60 patients suffering from MV regurgitation. For each 4D TEE acquisition, we extracted the 3D volume only at one time point with the MV in its closed state. This corresponded to mid-systole for General Electric HealthCare (GE HealthCare) and to end-systole for the Philips data. All 3D volumes were anonymized and exported to a Cartesian format with a fixed isotropic spacing of 0.5 mm. Furthermore, volumes acquired with the Philips system were spatially reoriented to match the coordinate system used in the GE HealthCare images.

Both the Philips and GE HealthCare datasets were randomly split into training, validation and test subsets, with a ratio of, respectively, 75%, 10% and 15%. Additionally, special care was taken to ensure that recordings from the same patient did not appear in more than one set split. All volumes were spatially cropped to $128 \times 128 \times 128$ voxels grids. A trained operator manually annotated the entire Philips dataset using the method described in [6] with 3D Slicer [16]. The MV was annotated using three classes by individually identifying the mitral annulus (MA), the anterior leaflet, and the posterior leaflet. The same trained operator also annotated the GE HealthCare test set (18 volumes), while the rest of the GE HealthCare dataset was left without annotations.

B. Self-training for Cross-Domain Segmentation

Borrowing the approach of [10], we implemented a semi-supervised teacher-student training strategy (Fig. 1), where the teacher model (T) was trained on the Philips data with supervision from manual annotations, and used to generate pseudo-labels for the unlabeled GE HealthCare data. To improve the pseudo-labels, T was based on ensemble modelling [17]. T consisted in the ensemble of three teacher models consisting each one in a different variation of a UNet-based network that was separately trained: a Residual UNet [18] (T_0), a Multi-Decoder Residual UNet [6] (T_1), and a SegResNet (T_2) [19]. The $argmax$ of their combined predictions over the GE HealthCare data was used to create the pseudo-labels.

The labeled and pseudo-labeled data were then combined to train the Student model (S), which was based on a SegResNet architecture due to its lightweight design and faster inference times, as well as it being the best of the T_j.

C. Pseudo-label enhancement

To mitigate the impact of unrealistic pseudo-labels during Student training, we applied strong data augmentation (SDA) and dropout to the GE HealthCare dataset. Following [10], we incorporated Gaussian blurring and Cutout transforms

in the standard data augmentation routine, which already included intensity scaling, random Gaussian noise, random rotation, random axis flip and random elastic deformation. SDA serves as a regularization technique, preventing S from over-fitting to incorrect pseudo-labels. A 0.1 dropout rate, determined through hyper-parameter search, was selected as it provided the best results in early experiments. Additionally, before training S, the pseudo-labels were processed using the reconstruction algorithm proposed in [20], which automatically restores a continuous and regular 3D profile for the MA in cases of under-segmentation or noisy prediction. A morphological closing image filter [21] was also applied to fill under-segmented regions in the interior or at the boundaries of the leaflet predictions.

D. Training and Evaluation Details

T_0, T_1, T_2, and S were trained over 500 epochs with a batch size of 2 or 1 on an Nvidia RTX A6000 with 48GB of memory using the Novograd optimizer with a 0.001 learning rate and Dice Focal loss. Their evaluation was carried out on two held-out test sets, one for the Philips and one for the GE HealthCare dataset of 20 volumes each, using the weights saved at the best validation epoch for the teachers, and the last epoch for the students. For the GE HealthCare dataset, we utilized the manual segmentations that were produced for the test set only, as described in Section III-A. Dice scores, Average Surface Distances (ASDs) and Hausdorff Distances 95% (HDFs 95%) were evaluated for the MV leaflets, while the curve-to-curve error was used for the MA as in [22], [23]. The evaluation metric for the MA was calculated following the application of the reconstruction algorithm from [6].

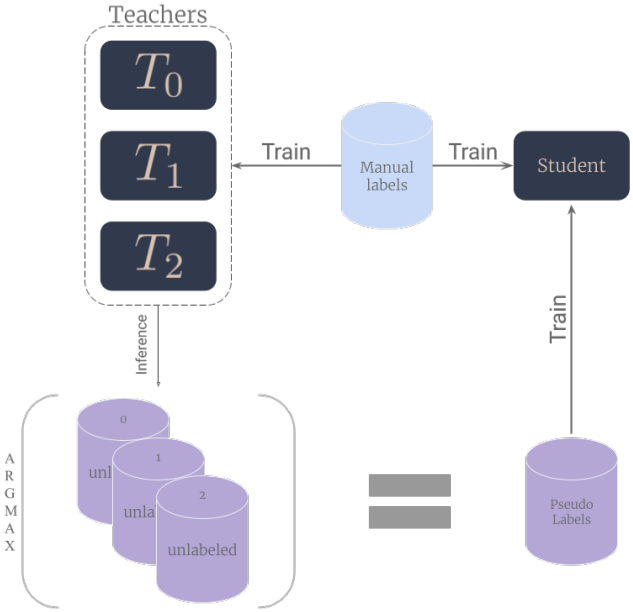

Fig. 1: Schematic view of our semi-supervised teacher-student training strategy

To assess the impact of SDA and post-processing on the pseudo-labels, we trained and evaluated the S network both with and without SDA, and with either refined or unrefined pseudo-labels.

IV. Results

A. Performance on Philips Dataset

Table I presents the performance of T_0, T_1, T_2, and S on the Philips test set. Among the individual teacher models, the SegResNet architecture (T_2) performed best, achieving a Dice score above 81% and a low HSD 95% of 1.53 ± 0.61 mm for the leaflets, and a curve-to-curve error of 1.66 ± 0.64 mm for the MA. As expected, the ensemble model T outperformed the individual teachers, achieving the highest Dice score ($82.12 \pm 4.80\%$) for the leaflets and the lowest curve-to-curve error (1.58 ± 0.63 mm) for the MA. Overall, the S model performed comparably to the ensemble T, with a Dice score of $82.20 \pm 4.62\%$ and a HSD 95% of around 1.6 mm for the leaflets, and a curve-to-curve error around 1.6 mm. Notably, the S model with SDA achieved the lowest ASD for the leaflets, with a value of 0.37 ± 0.07 mm. The use of SDA during its training did not visibly affect the performance of the S model on the Philips dataset. Fig. 2 shows qualitative segmentation results, which align with the quantitative results. Both the ensemble model T and the student model S demonstrated reliable MV segmentation, closely matching the GT. Even in the worst-case scenario (Fig. 2i to Fig. 2p), a complete MV reconstruction was achieved despite being spatially shifted relative to the GT (Figs. 2o and 2p), likely due to low image contrast, which made the task more challenging.

B. Performance on GE HealthCare Dataset

Table II presents the performance on the GE HealthCare test set. Similarly to the results on the Philips dataset, the SegResNet architecture (T_2) outperformed the 3D Residual UNet (T_0) and the Multi-Decoder UNet (T_1), yielding a Dice score above 69%, the lowest HSD 95% of all T_j, with 4.27 ± 2.63 mm for the leaflets, and a suitable curve-to-curve error of 5.55 ± 3.82 mm for the MA. The ensemble model T consistently improved performance over the individual teachers in terms of Dice score for the leaflets and distance metrics for both the leaflets and MA. Notably, compared to the results on the Philips dataset (Table I), the performance metrics were generally worse for the teacher models on the unseen GE HealthCare dataset. However, the S model effectively mitigated this performance drop, achieving the best Dice score ($71.32 \pm 6.95\%$) and the best HSD 95% (3.15 ± 1.61 mm) for the leaflets, and a curve-to-curve error of 4.51 ± 3.58 mm for the MA. As with the Philips dataset, the application of SDA only marginally impacted the performance of the S model. Also, the S model showed the best agreement with the GTs, providing more reliable segmentations for both the leaflets and the MA as compared to the teacher models (Fig. 3). It is worth noting that in the worst-case scenarios (the two bottom rows of Fig. 3) the S model struggled to fully reconstruct the leaflets and distinguish between the anterior and posterior

TABLE I: Evaluation of teachers and students network on the Philips dataset. For the students, the metrics are reported considering the refined pseudo-labels during training, with and without the application of SDA. The best values for each metric across all the models are highlighted in bold.

Metric	**Teacher**				**Student**	
	T_0	T_1	T_2	T	S w/o SDA	S w/ SDA
Dice score (%)	80.65 ± 7.60	80.71 ± 7.92	81.59 ± 5.09	82.12 ± 4.80	81.89 ± 4.44	**82.20 ± 4.62**
ASD (mm)	0.39 ± 0.07	0.38 ± 0.07	0.42 ± 0.11	0.42 ± 0.09	0.40 ± 0.08	**0.37 ± 0.07**
HSD 95% (mm)	1.86 ± 1.72	1.79 ± 1.86	**1.53 ± 0.61**	1.58 ± 1.02	1.63 ± 1.19	1.62 ± 1.24
Curve to curve (mm)	1.85 ± 1.16	1.80 ± 1.0.1	1.66 ± 0.64	**1.58 ± 0.63**	1.60 ± 1.12	1.63 ± 0.70

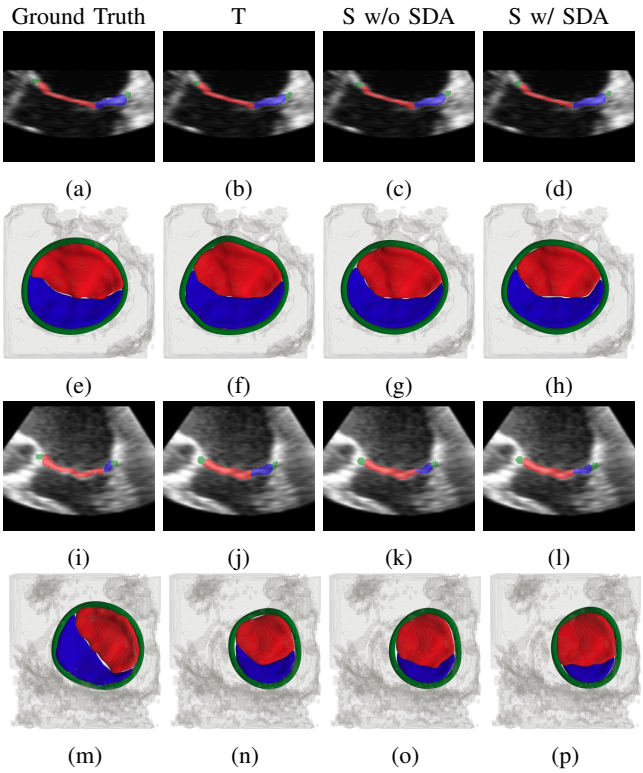

Fig. 2: Segmentation examples over the Philips dataset. The first two rows show 2D views and 3D renderings of our best case, and the bottom ones of the worst case, according to the Dice score. Color code: annulus is green, anterior leaflet is red, and posterior leaflet is blue.

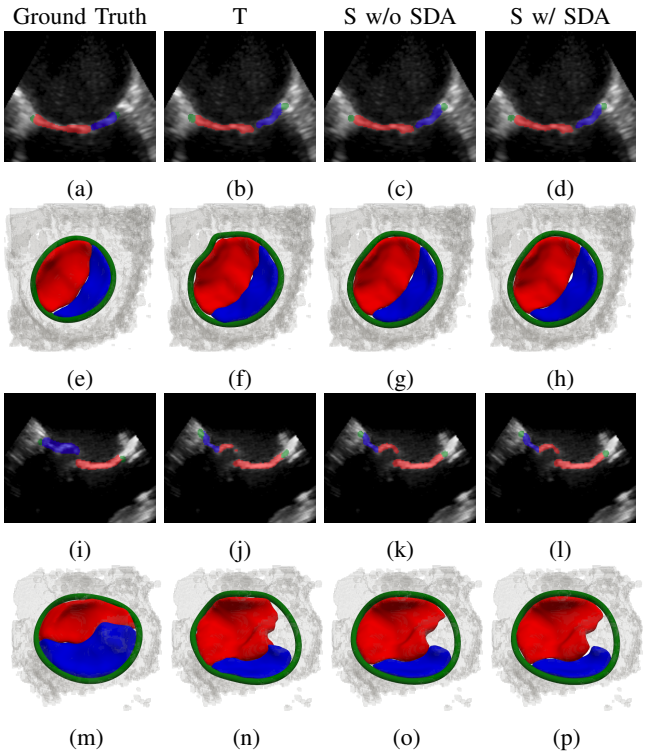

Fig. 3: Segmentation examples over the GE HealthCare dataset. The first two rows show 2D views and 3D renderings of our best case, and the bottom ones of the worst case, according to the Dice score. Color code: annulus is green, anterior leaflet is red, and posterior leaflet is blue.

leaflet, failing to identify the prolapse in the lateral portion of the posterior leaflet (Figs. 3i and 3m). Yet, the MA was fully reconstructed after applying the post-processing algorithm [6], showing good agreement with the GT (Figs. 3o and 3p).

C. Evaluation of pseudo-labels enhancement

Table III compares the performance of the S model on both test datasets when trained with either refined or unrefined pseudo-labels. Overall, the S model benefited from the use of the refinement algorithm [6], particularly on GE HealthCare data, showing a 2% point improvement in Dice score for the leaflets and a reduction in Curve-to-curve error for the MA. The impact of SDA was more pronounced when the pseudo-labels were not pre-processed with the reconstruction algorithm. Notably, the performance on the Philips dataset remained consistent, indicating that the refinement process did not degrade results on the original training domain.

V. Discussion

In this work, we applied a semi-supervised training strategy using pseudo-labeling to address the domain shift between US images from different vendors, focusing on automatic MV segmentation from 3D TEE. By leveraging a Teacher-Student framework, we combined labeled and pseudo-labeled data to train a robust segmentation model capable of generalizing across two US domains.

TABLE II: Evaluation of teachers and students network on the GE HealthCare dataset. For the students, the metrics are reported considering the refined pseudo-labels during training, with and without the application of SDA. The best values for each metric across all the models are highlighted in bold.

Metric	**Teacher**				**Student**	
	T_0	T_1	T_2	T	S w/o SDA	S w/ SDA
Dice score (%)	62.89 ± 8.44	62.01 ± 11.09	67.47 ± 6.92	69.38 ± 7.67	71.15 ± 6.49	**71.32 ± 6.95**
ASD (mm)	0.88 ± 0.21	1.23 ± 1.14	0.89 ± 0.26	0.88 ± 0.22	**0.82 ± 0.15**	0.83 ± 0.15
HSD 95% (mm)	4.43 ± 2.20	5.19 ± 4.09	4.27 ± 2.63	3.45 ± 1.60	**3.15 ± 1.61**	3.33 ± 1.65
Curve to curve (mm)	6.99 ± 4.65	7.58 ± 5.32	5.55 ± 3.82	**4.14 ± 3.66**	4.52 ± 3.74	4.51 ± 3.58

TABLE III: Comparison of pseudo label's refinement effect on training the student networks

Metric	**Unrefined**				**Refined**			
	Philips		GE HealthCare		Philips		GE HealthCare	
	S w/o SDA	S w/ SDA	S w/o SDA	S w/ SDA	S w/o SDA	S w/ SDA	S w/o SDA	S w/ SDA
Dice score (%)	82.10 ± 4.34	82.34 ± 4.43	68.47 ± 7.94	68.93 ± 6.98	81.89 ± 4.44	82.20 ± 4.62	71.15 ± 6.49	71.32 ± 6.95
ASD (mm)	0.37 ± 0.08	0.35 ± 0.07	0.86 ± 0.21	0.83 ± 0.18	0.40 ± 0.08	0.37 ± 0.07	0.82 ± 0.15	0.83 ± 0.15
HSD 95% (mm)	1.58 ± 1.00	1.71 ± 1.39	3.65 ± 1.98	3.38 ± 1.58	1.60 ± 1.12	1.62 ± 1.23	3.15 ± 1.61	3.33 ± 1.65
Curve to curve (mm)	1.70 ± 0.97	1.71 ± 1.12	4.57 ± 3.85	5.13 ± 4.02	1.72 ± 0.79	1.63 ± 0.70	4.52 ± 3.74	4.51 ± 3.58

Our approach utilized three state-of-the-art CNN architectures previously applied to MV segmentation from 3D TEE data [5]–[7]. Each model demonstrated unique strengths and limitations in this task. We employed ensemble modeling to mitigate these individual limitations while enforcing their strengths to ensure the creation of reliable pseudo-labels.

On the labelled Philips dataset, the individual teacher models (T_j, $j \in \{0, 1, 2\}$) achieved good metrics scores as compared to the GT (cf Table I). The ensemble model T further improved performance, particularly in terms of Dice score ($82.12 \pm 4.80\%$ for the leaflets) and distance error (1.58 ± 0.63 mm for the MA), confirming the advantage of ensemble methods in providing more accurate and robust predictions by averaging out biases [17]. The student model S performed comparably to, or even better than T, achieving the best Dice score ($82.20 \pm 4.62\%$) and the lowest ASD (0.37 ± 0.07 mm) for the leaflets, resulting in more reliable MV segmentation masks (Figs. 2c, 2d, 2g and 2h). The use of pseudo-labels generated by the ensemble T on the unlabeled GE HealthCare dataset did not negatively affect the performance of the S model on the Philips dataset. Instead, the S model appeared to benefit from the larger, combined training dataset, which likely helped it learn better MV features. Although direct comparisons cannot be made due to differences in datasets, the results are comparable to previous works on single-vendor MV segmentation [6]–[8], which report Dice scores ranging between 81% and 87.7%, ASD values between 0.32 mm and 0.92 mm, and HSD 95% values between 1.86 mm and 1.99 mm. Similarly, the curve-to-curve errors equal to 1.58 and 1.60 mm obtained by the ensemble model T and the student model S for the MA are comparable to those from previous works reporting curve-to-curve errors equal to 2.04 mm [22] and 1.96 mm [23].

On the unlabeled GE HealthCare dataset, the teacher models showed a significant drop in performance, as expected. The ensemble model T, again, provided the best results among them, with Dice score just below 70%, ASD of 0.88 ± 0.22 mm and HSD 95% of 3.45 ± 1.60 mm for the leaflets, as well as curve-to-curve error of 4.14 ± 3.66 mm for the MA. On the other hand, the S model outperformed the ensemble T, achieving a better Dice score ($71.32 \pm 6.95\%$), ASD (0.82 ± 0.15 mm) and HSD 95% (3.15 ± 1.61 mm) for the leaflets, resulting in more reliable MV segmentation masks (Figs. 3c, 3d, 3g and 3h). S model's improved performance on the unseen GE HealthCare domain demonstrates the effectiveness of our strategy, which leverages additional training examples through pseudo-labeling. The ensemble model T generated realistic and reliable pseudo-labels, outperforming individual teachers and enhancing generalization across domains, as confirmed by the results on the GE HealthCare test set. Any residual uncertainty in the pseudo-labels was likely mitigated by incorporating SDA in the student training. Evaluating the quality of pseudo-labels and the impact of label noise on deep neural network training is a critical aspect of this approach. In this study, we assessed these factors by using a small subset of manually annotated volumes from the GE HealthCare dataset. Alternative methods for evaluating the impact of label noise exist [24], including approaches that do not require a noise-free subset or prior knowledge of the noise level in the labels [25]. However, these analyses were considered beyond the scope of this work. The favourable results achieved on the GE HealthCare dataset, particularly through ensemble modeling, demonstrate that our strategy effectively addresses challenges associated with pseudo-label noise and enhances cross-domain segmentation performance.

The student performance on unseen GE HealthCare dataset benefited from the proposed pseudo-label refinement, which efficiently reduced the impact of possible unrealistic pseudo-labels during student training while maintaining performance on the Philips dataset, as shown in Table III. Unexpectedly, the role of SDA, which mitigates the impact of incorrect training examples in student training, as described in [10], was

negligible when refined pseudo-labels were used. This further confirms the effectiveness of the proposed pseudo-label refinement, which ensures high-quality training examples, reducing the uncertainty to mitigate and helping the model learn robust features from pseudo-labeled data.

However, the performance on the GE HealthCare dataset was overall worse than compared to the Philips dataset and previous works on single-vendor MV segmentation. This discrepancy may be attributed to the generally higher variability of the available GE HealthCare dataset which is representative of a wider range of MV conditions: healthy and affected by primary or secondary mitral regurgitation, or by mitral stenosis, whereas the Philips dataset is only focused on mitral regurgitation. This means that in addition to a domain shift between the US vendors, there is also a distribution shift in the represented MV shapes. Our method addressed the US domain shift effectively. However, variations in disease distribution remain a challenge to be tackled, since different diseases or even different aetiologies of the same disease can lead to different deformations in MV anatomy, e.g., MA size and shape, and leaflet thickness and extent, and in MV closed configuration, which is characterized by billowing leaflets in primary regurgitation and by tethered leaflets in secondary regurgitation. Anonymizing the datasets prevented linking specific disease conditions to TEE acquisitions, limiting disease-specific analysis and optimization.

Our semi-supervised approach reduces annotation workload and tackles domain shift challenges in 3D TEE MV segmentation. Future studies could explore complementary weakly supervised training methods to enhance these results further. For example, a pyramid network approach was introduced in [26] to learn image segmentation tasks from both labeled and unlabeled data, integrating uncertainty estimation within a pyramid consistency framework to better leverage the impact of unlabeled data. More recently, a hybrid CNN-transformer approach was proposed in [27] to perform efficient segmentation leveraging scribble annotations. These alternative weakly supervised training methods are well-suited for addressing domain shift challenges and should be considered for future comparisons.

Additionally, a direct quantitative comparison with domain adaptation models [14], [15] was not feasible due to the challenges of training such networks on 3D US images. It is reasonable to assume that the proposed self-training strategy for efficient cross-domain MV segmentation from 3D TEE is more computationally attainable than using a 3D GAN or a diffusion model. The proposed method required only a marginal increase in computational cost as compared to a traditional fully-supervised learning approach, while ensuring strong generalization across different domains and delivering state-of-the-art performance for MV segmentation.

Finally, the proposed strategy was evaluated using datasets from the two biggest cardiovascular ultrasound vendors. GE HealthCare and Philips accounted for 48% of the ultrasound equipment in 2023, but to further demonstrate its potential and applicability in real-world scenarios, future work should involve a broader comparison across additional US vendors, as well as a more comprehensive representation of MV diseases.

VI. Conclusion

We proposed a vendor-agnostic and automatic segmentation tool for 3D TEE using a teacher-student training strategy. We showed that this strategy allows for domain shift between two of the main US manufacturers, as well as improves inference over the initial US domain. This is a significant step towards vendor-agnostic segmentation of 3D TEE.

Acknowledgment

This work was made possible thanks to the financial support of the Norwegian University of Science and Technology's Department of Computer Science, the Department of Electronics, Information and Bioengineering of Politecnico di Milano and the Centre for Innovative Ultrasound Solutions, as well as the European Union's Horizon 2020 research and innovation program, under the project ARTERY, grant agreement No. 101017140 and the Centre for Innovative Ultrasound Solutions.

We would also like to thank General Electric HealthCare and the Department of Circulation and Medical Imaging of the Norwegian University of Science and Technology at St. Olavs Hospital along with the Unit of Cardiovascular Imaging and the Cardiac Surgery Department of San Raffaele Hospital for allowing us to use their data.

References

[1] C. W. Tsao, A. W. Aday, Z. I. Almarzooq, A. Alonso, A. Z. Beaton, M. S. Bittencourt, A. K. Boehme, A. E. Buxton, A. P. Carson, Y. Commodore-Mensah, M. S. Elkind, K. R. Evenson, C. Eze-Nliam, J. F. Ferguson, G. Generoso, J. E. Ho, R. Kalani, S. S. Khan, B. M. Kissela, K. L. Knutson, D. A. Levine, T. T. Lewis, J. Liu, M. S. Loop, J. Ma, M. E. Mussolino, S. D. Navaneethan, A. M. Perak, R. Poudel, M. Rezk-Hanna, G. A. Roth, E. B. Schroeder, S. H. Shah, E. L. Thacker, L. B. VanWagner, S. S. Virani, J. H. Voecks, N.-Y. Wang, K. Yaffe, S. S. Martin, and on behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee, "Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association," *Circulation*, vol. 145, pp. e153–e639, Feb. 2022. Publisher: American Heart Association.

[2] A. Vahanian, F. Beyersdorf, F. Praz, M. Milojevic, S. Baldus, J. Bauersachs, D. Capodanno, L. Conradi, M. De Bonis, R. De Paulis, V. Delgado, N. Freemantle, M. Gilard, K. H. Haugaa, A. Jeppsson, P. Jüni, L. Pierard, B. D. Prendergast, J. R. Sádaba, C. Tribouilloy, W. Wojakowski, ESC/EACTS Scientific Document Group, and ESC National Cardiac Societies, "2021 ESC/EACTS Guidelines for the management of valvular heart disease: Developed by the Task Force for the management of valvular heart disease of the European Society of Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS)," *European Heart Journal*, vol. 43, pp. 561–632, Feb. 2022.

[3] N. Thomas, B. Unsworth, E. A. Ferenczi, J. E. Davies, J. Mayet, and D. P. Francis, "Intraobserver variability in grading severity of repeated identical cases of mitral regurgitation," *American Heart Journal*, vol. 156, pp. 1089–1094, Dec. 2008.

[4] E. Costa, N. Martins, M. S. Sultan, D. Veiga, M. Ferreira, S. Mattos, and M. Coimbra, "Mitral Valve Leaflets Segmentation in Echocardiography using Convolutional Neural Networks," in *2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG)*, pp. 1–4, Feb. 2019.

[5] P. Carnahan, J. Moore, D. Bainbridge, M. Eskandari, E. C. S. Chen, and T. M. Peters, "DeepMitral: Fully Automatic 3D Echocardiography Segmentation for Patient Specific Mitral Valve Modelling," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* (M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, eds.), Lecture Notes in Computer Science, (Cham), pp. 459–468, Springer International Publishing, 2021. Philips images.

[6] R. Munafò, S. Saitta, G. Ingallina, P. Denti, F. Maisano, E. Agricola, A. Redaelli, and E. Votta, "A Deep Learning-Based Fully Automated Pipeline for Regurgitant Mitral Valve Anatomy Analysis From 3D Echocardiography," *IEEE Access*, vol. 12, pp. 5295–5308, 2024. Conference Name: IEEE Access.

[7] P. Carnahan, A. Bharucha, M. Eskandari, E. C. S. Chen, and T. M. Peters, "Segmentation of the Mitral Valve from 3D Transesophageal Echocardiography," Apr. 2023. Publisher: Zenodo.

[8] A. H. Aly, P. Khandelwal, A. H. Aly, T. Kawashima, K. Mori, Y. Saito, J. Hung, J. H. Gorman, A. M. Pouch, R. C. Gorman, and P. A. Yushkevich, "Fully Automated 3D Segmentation and Diffeomorphic Medial Modeling of the Left Ventricle Mitral Valve Complex in Ischemic Mitral Regurgitation," *Medical Image Analysis*, vol. 80, p. 102513, Aug. 2022.

[9] J. Chen, H. Li, G. He, F. Yao, L. Lai, J. Yao, and L. Xie, "Automatic 3D mitral valve leaflet segmentation and validation of quantitative measurement," *Biomedical Signal Processing and Control*, vol. 79, p. 104166, Jan. 2023. Philips images.

[10] R. Munafò, S. Saitta, D. Tondi, G. Ingallina, P. Denti, F. Maiasano, E. Agricola, and E. Votta, "Automatic 4D mitral valve segmentation from transesophageal echocardiography: a semi-supervised learning approach," *TechRxiv*.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, vol. 27, Curran Associates, Inc., 2014.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models,"

[13] W. Yan, Y. Wang, M. Xia, and Q. Tao, "Edge-Guided Output Adaptor: Highly Efficient Adaptation Module for Cross-Vendor Medical Image Segmentation," *IEEE Signal Processing Letters*, vol. 26, pp. 1593–1597, Nov. 2019. Conference Name: IEEE Signal Processing Letters.

[14] T. Chen, M. Xia, Y. Huang, J. Jiao, and Y. Wang, "Cross-Domain Echocardiography Segmentation with Multi-Space Joint Adaptation," *Sensors*, vol. 23, p. 1479, Jan. 2023. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[15] C. Tiago, S. R. Snare, J. Šprem, and K. McLeod, "A Domain Translation Framework With an Adversarial Denoising Diffusion Model to Generate Synthetic Datasets of Echocardiography Images," *IEEE Access*, vol. 11, pp. 17594–17602, 2023. Conference Name: IEEE Access.

[16] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J. V. Miller, S. Pieper, and R. Kikinis, "3D Slicer as an image computing platform for the Quantitative Imaging Network," *Magnetic Resonance Imaging*, vol. 30, pp. 1323–1341, Nov. 2012.

[17] H. Li, G. Jiang, J. Zhang, R. Wang, Z. Wang, W.-S. Zheng, and B. Menze, "Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images," *NeuroImage*, vol. 183, pp. 650–665, Dec. 2018.

[18] E. Kerfoot, J. Clough, I. Oksuz, J. Lee, A. P. King, and J. A. Schnabel, "Left-Ventricle Quantification Using Residual U-Net," in *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges* (M. Pop, M. Sermesant, J. Zhao, S. Li, K. McLeod, A. Young, K. Rhode, and T. Mansi, eds.), Lecture Notes in Computer Science, (Cham), pp. 371–380, Springer International Publishing, 2019.

[19] M. M. R. Siddique, D. Yang, Y. He, D. Xu, and A. Myronenko, "Automated ischemic stroke lesion segmentation from 3D MRI," Sept. 2022. arXiv:2209.09546 [cs, eess].

[20] R. J. Schneider, D. P. Perrin, N. V. Vasilyev, G. R. Marx, P. J. del Nido, and R. D. Howe, "Mitral annulus segmentation from 3D ultrasound using graph cuts," *IEEE transactions on medical imaging*, vol. 29, pp. 1676–1687, Sept. 2010.

[21] G. Lehmann, "Binary morphological closing and opening image filters," *The Insight Journal*, Nov. 2005.

[22] B. S. Andreassen, F. Veronesi, O. Gerard, A. H. S. Solberg, and E. Samset, "Mitral Annulus Segmentation Using Deep Learning in 3-D Transesophageal Echocardiography," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 994–1003, Apr. 2020. Conference Name: IEEE Journal of Biomedical and Health Informatics.

[23] B. S. Andreassen, D. Völgyes, E. Samset, and A. H. S. Solberg, "Mitral Annulus Segmentation and Anatomical Orientation Detection in TEE Images Using Periodic 3D CNN," *IEEE Access*, vol. 10, pp. 51472–51486, 2022. Conference Name: IEEE Access.

[24] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from Noisy Labels with Deep Neural Networks: A Survey," Mar. 2022. arXiv:2007.08199 [cs].

[25] B. D. de Vos, G. E. Jansen, and I. Išgum, "Stochastic co-teaching for training neural networks with unknown levels of label noise," *Scientific Reports*, vol. 13, p. 16875, Oct. 2023. Publisher: Nature Publishing Group.

[26] X. Luo, G. Wang, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, D. N. Metaxas, and S. Zhang, "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," *Medical Image Analysis*, vol. 80, p. 102517, Aug. 2022.

[27] Z. Li, Y. Zheng, D. Shan, S. Yang, Q. Li, B. Wang, Y. Zhang, Q. Hong, and D. Shen, "ScribFormer: Transformer Makes CNN Work Better for Scribble-Based Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 43, pp. 2254–2265, June 2024. Conference Name: IEEE Transactions on Medical Imaging.