Medical & Biological Engineering & Computing
https://doi.org/10.1007/s11517-024-03275-w

ORIGINAL ARTICLE

# Automatic 4D mitral valve segmentation from transesophageal echocardiography: a semi-supervised learning approach

Riccardo Munafò[1] · Simone Saitta[1,2,6] · Davide Tondi[1] · Giacomo Ingallina[3] · Paolo Denti[4] · Francesco Maisano[4] · Eustachio Agricola[3,5] · Emiliano Votta[1]

Received: 20 September 2024 / Accepted: 18 December 2024


**Abstract**
Performing automatic and standardized 4D TEE segmentation and mitral valve analysis is challenging due to the limitations of echocardiography and the scarcity of manually annotated 4D images. This work proposes a semi-supervised training strategy using pseudo labelling for MV segmentation in 4D TEE; it employs a Teacher-Student framework to ensure reliable pseudo-label generation. 120 4D TEE recordings from 60 candidates for MV repair are used. The Teacher model, an ensemble of three convolutional neural networks, is trained on end-systole and end-diastole frames and is used to generate MV pseudo-segmentations on intermediate frames of the cardiac cycle. The pseudo-annotated frames augment the Student model's training set, improving segmentation accuracy and temporal consistency. The Student outperforms individual Teachers, achieving a Dice score of 0.82, an average surface distance of 0.37 mm, and a 95% Hausdorff distance of 1.72 mm for MV leaflets. The Student model demonstrates reliable frame-by-frame MV segmentation, accurately capturing leaflet morphology and dynamics throughout the cardiac cycle, with a significant reduction in inference time compared to the ensemble. This approach greatly reduces manual annotation workload and ensures reliable, repeatable, and time-efficient MV analysis. Our method holds strong potential to enhance the precision and efficiency of MV diagnostics and treatment planning in clinical settings.

**Keywords** Three-dimensional transesophageal echocardiography · Mitral regurgitation · Automatic segmentation · Convolutional neural network · Semi-supervised learning · Mitral valve

## 1 Introduction

The mitral valve (MV) is a dynamic structure between the left ventricle and atrium, consisting of two leaflets anchored to a saddle-shaped annulus and connected to the myocardium by *chordae tendineae* that attach to the papillary muscles. In physiological MVs, during diastole the leaflets open wide and the annulus dilates to allow blood flow from the left atrium to the left ventricle. In systole, the leaflets close tightly and the annulus contracts to prevent backflow into the atrium, ensuring valve competence and optimal leaflet coaptation. Disruption of any MV component can lead to mitral regurgitation (MR) [1], the most widespread valvular pathology [2]. Echocardiography is the primary imaging modality for visualizing and assessing MV anatomy and function [3]. Specifically, transesophageal echocardiography (TEE) provides high-quality imaging of cardiac structures in 2D and 3D, enabling the dynamic evaluation of these structures in real-time. TEE is also adopted to monitor MV repair [7], including transcatheter edge-to-edge repair (TEER). Quantitative assessment of mitral annulus dynamics, leaflet mobility, and excursion is crucial in evaluating MR and characterizing its various etiologies [4, 5]. Furthermore, a dynamic quantification of MV morphology and kinematics can inform computational models, including finite element simulations that replicate patient-specific MV movement and mechanical properties [10, 11]. The quantification of the MV apparatus from 4D TEE, possibly distinguishing between its

✉ Riccardo Munafò
riccardo.munafo@polimi.it

1 Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

2 Department of Biomedical Engineering and Physics, Amsterdam UMC, Amsterdam, The Netherlands

3 Unit of Cardiovascular Imaging, IRCCS San Raffaele Hospital, Milan, Italy

4 Cardiac Surgery Department, IRCCS San Raffaele Hospital, Milan, Italy

5 Vita-Salute San Raffaele University, Milan, Italy

6 Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

Published online: 11 January 2025

substructures, requires pixel-wise annotation. If not highly automated, this process is operator-dependent [12], as well as labor-intensive and time-consuming, often exceeding the time constraints of a standard examination, even for expert sonographers [13]. Automatic and standardized methods for 4D TEE annotation and MV morphology and motion quantification can reduce the sonographers' workload and improve the analysis reliability and repeatability.

Automatic methods for MV segmentation and tracking from 4D TEE are challenging due to the well-known limitations of echocardiography, including motion artifacts, poor-image quality, and scarce availability of manually annotated 4D images. Convolutional neural networks (CNNs) have shown promising performance for supervised MV segmentation from echocardiography images. Costa et al. [20] implemented a 2D CNN for automatic MV leaflets segmentation from 2D transthoracic echocardiography (TTE) data. Carnahan et al. [21] and Aly et al. [22] proposed automatic 3D MV segmentation from 3D TEE images using a 3D Residual UNet and nnUNet, respectively. Chen et al. [23] proposed a two-step solution for 3D MV segmentation in systolic and diastolic configurations, initializing a nnUNet with pretrained weights of a classifier for the MV open and closed states. Munafò et al. [33] developed a Multi-Decoder Residual UNet for multi-structure MV segmentation at end-systole (ES), enabling separate identification of the mitral annulus, anterior and posterior leaflets from 3D TEE images. Yet, none of the mentioned studies achieved the frame-by-frame quantification of MV anatomy in 4D TEE images. Indeed, so far this task has been performed either on 3D TEE imaging but with a focus only on annulus segmentation [25, 26] or tracking [27], or on annulus and leaflets but in 2D imaging, as in [24], who proposed an effective UNet with attention gates for segmenting 2D TTE images. Despite the advancements in neural network architectures, there is currently no method allowing for the full 4D MV segmentation within the short times required to meet the strict clinical demands of diagnostic echocardiographic screening, where large amounts of dataset must be analyzed, and intraprocedural guidance during percutaneous interventions, where real-time or almost real-time annotation of images is needed to provide operators with timely visual information. A significant bottleneck in the development of such a method is the labor-intensive manual annotation process needed to generate the extensive datasets required for the supervised training of CNNs.

This work addresses this gap by introducing a semi-supervised training strategy that reduces the need for manual annotations without compromising segmentation accuracy. Our approach leverages a semi-supervised training strategy using pseudo-labeling, which has never been applied in the context of 4D TEE MV segmentation. We use manual annotations of only the ES and end-diastole (ED) frames to achieve automatic segmentation of the MV throughout the cardiac cycle. Additionally, we propose an ensemble strategy to generate high-confidence pseudo-labels and evaluate the impact of specific data augmentation previously used in self-training strategies to ensure high consistency on the unlabeled dataset.

## 2 Methods

### 2.1 Dataset collection and annotation

Intraprocedural examinations were collected from 60 candidates for MV TEER using the MitraClip system (Abbott) at IRCCS San Raffaele Hospital (Milan, Italy). The cohort consisted of 21 females and 39 males, with ages ranging from 50 and 95 years. Acquisitions were performed with a Philips EPIQ CVx scanner and an X8-2T transducer (Philips, Andover, MA, USA). Image data were anonymized and exported in Cartesian format using QLab software (Philips, Andover, MA, USA). The study was approved by the local ethics committee. Midesophageal 4D TEE acquisitions, zoomed in on the MV, were selected from the collected examinations to construct the dataset. For some patients, multiple acquisitions were available and included in the dataset. Acquisitions with high shadowing or those that did not entirely capture the MV within the field of view of the ultrasound beam were excluded. The final dataset counted $n = 120$ 4D TEE recordings with a mean voxel spacing of $0.37\ \text{mm} \times 0.55\ \text{mm} \times 0.24\ \text{mm}$ and a mean frame rate of 22 per cardiac cycle. This dataset is a slightly augmented version of the dataset used in our previous work [33], with a small increase in the number of patients considered. Additionally, manual 3D annotations of the MV at the ES frame from [33] were reused in this work and expanded to include MV annotations at ED. ES and ED were defined as the frame before the aortic valve closure and the frame before MV closure, respectively. ED volumes were annotated by three independent and experienced operators following the same protocol described in [33] and using 3D Slicer [43]. Multi-class annotations of the MV were produced by separately identifying the mitral annulus, the anterior leaflet, and the posterior leaflet.

### 2.2 Semi-supervised segmentation strategy

Semi-supervised semantic segmentation aims to generalize using a combined set of voxel-wise labeled images $D^l = \{(x_i, y_i)\}_{i=1}^{M}$ and unlabeled images $D^u = \{(x_i)\}_{i=1}^{N}$, with $N$ usually larger than $M$. In this work, the images are from $n$ 4D TEE recordings of the MV covering a full cardiac cycle, where at each frame a volumetric image is available. $D^l$ consists of the 3D images at ES and ED frames extracted from each recording, along with the corresponding manually obtained 3D annotations of MV substructures. $D^u$ comprises the images at three intermediate frames evenly sampled over

the remaining frames of the cardiac cycle (Fig. 1). Overall, $D^l$ and $D^u$ contain a total of $M = 2n$ and $N = 3n$ images, respectively. Both $D^l$ and $D^u$ are used complementary in a self-training learning strategy to train CNN models for multi-class segmentation of the MV.

### 2.3 Self-training approach for MV segmentation

For the formulation of the self-training approach, we leverage terminology specific to the Teacher-Student framework. In self-training, the Teacher model $T$ is first trained on labeled data $D^l$ and then used to assign pseudo-labels to the unlabeled data point $x_i \in D^u$ [31]. The combination of the labeled and pseudo-labeled sets is used to train the Student model $S$. Previous formulations of self-training approaches may differ depending on the heuristics used to select the proxy-labeled examples to include in the labeled set. In this work, we propose to employ ensemble modeling to ensure reliable pseudo-labels. Specifically, we consider three different CNN architectures for image segmentation, which have been previously adopted for research studies [21, 33] or challenges [36] related to MV segmentation from 3D TEE images. These architectures are separately trained on $D^l$ using the same training scheme. The ensemble of these three models is then used to predict the pseudo-labels by averaging the probability maps from each model. Combining multiple networks into an ensemble typically improves results over using a single CNN [34]. Figure 2 shows a complete overview of the adopted self-training approach. Similarly to [32] and according to Fig. 2, the proposed self-training approach can be summarized in the following three steps:

1. **Teacher Training**: train three CNN models, $T_1$, $T_2$, $T_3$ on $D^l$.
2. **Pseudo Labelling**:
   (a) Use $T_1$, $T_2$, $T_3$ to make predictions on $D^u$.
   (b) Compute the average predicted probability map across the three models as $\hat{P}(x_i) = \frac{1}{3}\sum_j^3 P_j(x_i)$ with $x_i \in D^u$ and assign the pseudo-labels $\tilde{y}_i = \arg\max \hat{P}(x_i)$ by taking the argmax of $\hat{P}(x_i)$. Hence, in this formulation, the actual Teacher $T$ consists of the ensemble of $T_j$, with $j = 1, 2, 3$.
3. **Student Training**:
   (a) Generate the dataset to train the Student $S$. This is the combination of $D^l$ with $\tilde{D}^u$, where $\tilde{D}^u = \{(x_i, \tilde{y}_i)\}_{i=1}^N$ is defined by combining each data $x_i \in D^u$ with the corresponding pseudo-label $\tilde{y}_i$.
   (b) Train a randomly initialized Student $S$ on the combined $D^l$ and $\tilde{D}^u$.

In the described process, all datasets belong to the training set. Generalization performance will be evaluated on a held-out test set.

### 2.4 Data augmentation on unlabeled images

To increase the robustness against less reliable pseudo-labels and ensure balanced performance across $D^l$ and $D^u$, strong data augmentation (SDA) is applied to unlabeled images during $S$ training. This approach acts as a regularization technique, presenting a more challenging optimization target and preventing overfitting to less reliable pseudo-labels. Inspired by [32], Gaussian blurring and CutOut transforms [35] are included in the basic data augmentation routine. Gaussian blurring is an intensity-based transform that reduces image details while preserving edges by convolving the image with a Gaussian function. For this transform, sigma values are randomly selected between 0.5 and 1. CutOut is a spatial transform that randomly removes patches of a specific size from the images. For this transform, homogeneous patches with dimensions randomly selected between 2 and 20% of the image's input dimensions are used. The application of both transforms on unlabeled images has been shown to be effective in self-training approaches, as demonstrated in [32].

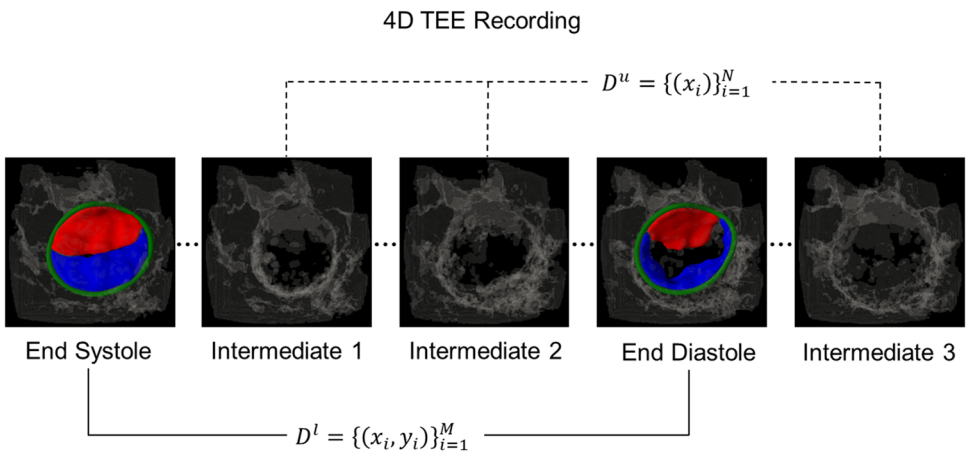


**Fig. 1** 3D renderings of TEE frames extracted from each 4D TEE recording, where each frame consists of a 3D image volume. The ES and the ED frames, along with the corresponding 3D MV annotations, compose $D^l$. Three intermediate frames, belonging to the diastole and systole phases and picked to uniformly sample the cardiac cycle, compose $D^u$

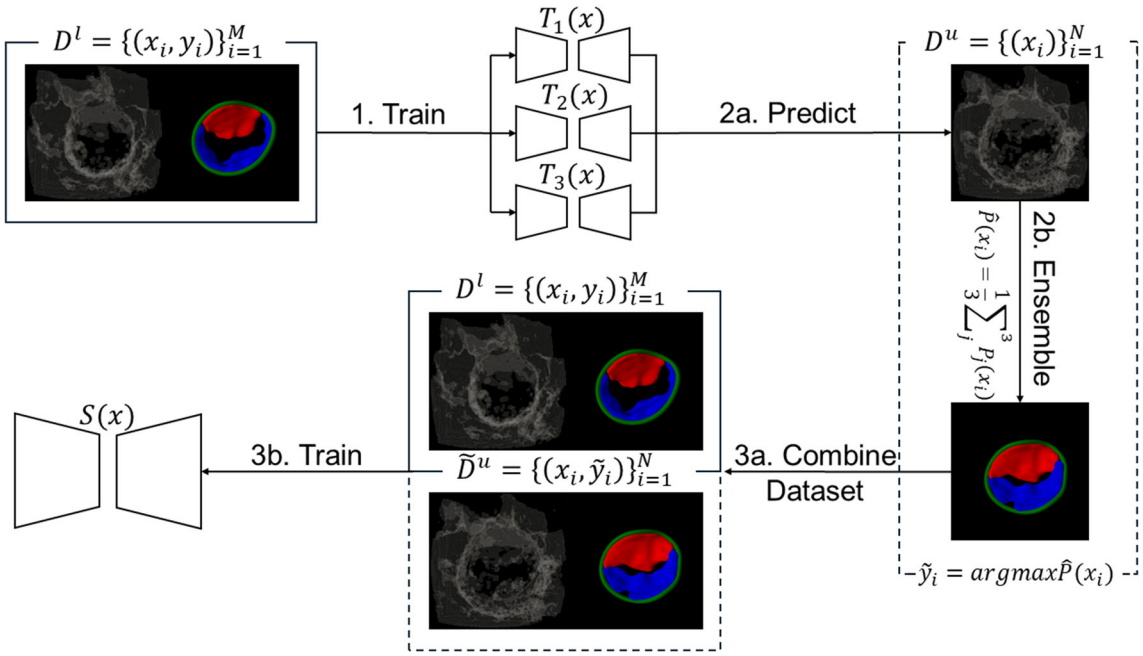


**Fig. 2** Overview of the proposed self-training strategy. Step 1, **Teacher Training**. $T_1$, $T_2$, and $T_3$ are trained on $D^l$, i.e., on 3D TEE images acquired at ES and ED. Step 2, **Pseudo Labelling**. (a) $T_1, T_2, T_3$ are used to make predictions on $D^u$, i.e., volumetric 3D TEE images acquired at intermediate systolic and diastolic frames of the cardiac cycle. Then, (b) pseudo-labels $\tilde{y}_i$ are generated by ensembling the predictions from the three models. Step 3, **Student Training**. (a) $D^l$ and $\tilde{D}^u = \{(x_i, \tilde{y}_i)\}_{i=1}^N$ are combined, and (b) their combination is used to train the Student $S$ model

To balance the impact of labeled and unlabeled data during $S$ training, $D^l$ is oversampled to match the size of $D^u$. Specifically, each mini-batch is constructed to include an equal number of labeled and unlabeled images. This approach allows for balanced sampling of mini-batches from both datasets, ensuring that $D^l$ and $D^u$ equally contribute to each weight update during $S$ training, thereby preventing bias towards the larger unlabeled dataset.

## 2.5 Neural network architectures

Three architectures are considered for the Teacher models: Residual UNet, Multi-Decoder Residual UNet, and SegResNet. Residual UNet was successfully applied in [21] for MV 3D segmentation from 3D TEE images at ED. Multi-Decoder Residual UNet was proposed in our previous work [33] for efficient multi-structure MV 3D segmentation from 3D TEE images at ES. SegResNet won the MICCAI 2023 challenge on MV segmentation from 3D TEE [36]. These architectures share common beneficial features for MV segmentation from 3D TEE images but differ in complexity and in the number of trainable parameters.

### 2.5.1 Residual UNet

The Residual UNet is an encoder-decoder-based CNN architecture built upon the UNet framework [37]. It consists of five resolution levels, each one defined using residual units. In the encoder path, residual units include two convolutional blocks and an identity mapping. In the decoder path, the residual units include one convolutional block and an identity mapping. A transposed convolutional layer is used for upsampling. Each convolutional block includes a 3×3 convolutional layer with a stride of 2 in the encoding path, a batch normalization layer, and a parametric rectifying linear (PReLU) unit. This model has approximately 20 million trainable parameters.

### 2.5.2 Multi-Decoder Residual UNet

The Multi-Decoder Residual UNet is an extension of the Residual UNet with multiple decoder branches [33]. The backbone architecture is borrowed from the Residual UNet, but an equal number of convolutional blocks are used in the residual units of the encoder and decoder branches, resulting in a more symmetric structure. The number of decoder branches is extended to three, one for each detected class (mitral annulus, anterior leaflet, and posterior leaflet). This model has around 40 million trainable parameters.

### 2.5.3 SegResNet

The SegResNet is a UNet-based architecture with an asymmetrically larger encoder to extract image features and a smaller decoder to reconstruct the segmentation mask [38].

The encoder part uses ResNet blocks, which differ from previously described residual units by incorporating group normalization layers and ReLU units. Four stages of downsampling are used, each stage has 1, 2, 2, and 4 ResNet blocks, respectively. In the decoder part, a single ResNet block per spatial level is used, and 3D bilinear upsampling restores the original input dimension. SegResNet is the lightest architecture considered, with around 18 million trainable parameters.

#### 2.5.4 Teacher and Student architecture

A Residual UNet, Multi-Decoder Residual UNet, and SegResNet are used individually to construct $T_1$, $T_2$, and $T_3$, respectively. The prediction of these models is then ensembled to form the Teacher $T$. The ensemble approach exploits the strengths of each architecture by combining their predictions to generate more reliable and consistent pseudo-labels. For the Student $S$, SegResNet is chosen because of its lightweight architecture, which ensures shorter inference times.

### 2.6 Training and implementation details

The Residual UNet, the Multi-decoder UNet, and the SegResNet are separately trained on $D^l$, using 70% of the images for training, 10% for validation, and 20% for testing. Data splitting is performed at the patient level to ensure that acquisitions from the same patient are not present in both training and evaluation sets. Images are resampled with a homogeneous spacing of 0.5 mm along each direction. Patches of $128 \times 128 \times 128$ voxels are randomly extracted from the input 3D image and passed to the model during the training phase. Data augmentation is performed *on-the-fly*, including intensity transforms such as intensity scaling and random Gaussian noise, as well as spatial transforms such as random rotation, random axis flip, and random elastic deformation. No transformation is applied during validation, whereas the largest connected component is extracted from the predicted segmentation during testing. Each model is trained over 500 epochs with a batch size of 2, using the Novograd optimizer [39] with a learning rate of 0.001. A weighted combination of Dice and Focal losses [40], with weights of 0.6 and 0.4, respectively, is minimized during the training phase. The saved model at the last epoch is used for inference on the test set.

After creating pseudo-labels $\tilde{y}_i$ in $D^u$, for the Student model, a new randomly initialized SegResNet architecture is trained on the combined set $D^l$ and $D^u$. The Student $S$ is trained using the previously described data splitting and training scheme, with the addition of SDA on $D^u$. The basic data augmentation routine is augmented with random Gaussian blurring and random CutOut for the unlabeled dataset. All models were implemented using a PyTorch framework [41] and the Python-based medical imaging library MONAI [42]. Training and inference of the model were performed on an NVIDIA A4000 with 16 GB of memory.

## 3 Evaluation

The test set consists of 16 4D TEE recordings. These are used to quantitatively evaluate if the proposed self-learning strategy fulfills two aims.

First, augmenting the training dataset for the Student $S$ while maintaining high accuracy on the labeled data, demonstrating that pseudo-labels do not degrade performance. To this aim, each individual Teacher model $T_j$, the ensemble Teacher $T$, and the Student $S$ are evaluated separately on $D^l$: the respective predictions $\hat{y}_i$ are compared vs. ground truth (GT) annotations $y_i$ by an expert operator for each $x_i \in D^l$ (Fig. 3). Standard segmentation metrics such as Dice score, average surface distance (ASD, in mm), and 95% Hausdorff distance (95%HD, in mm) are computed for the mitral leaflets. For the annulus, coordinate prediction metrics adopted in previous works [25, 26], referred to as curve-to-curve distance error (in mm) and in-plane error (in mm), are used. Curve-to-curve distance error represents the average shortest error between the predicted and the GT annulus 3D curve. In-plane error is the Euclidean distance between predicted and GT annulus points sampled from each rotating long-axis view around the left ventricle's main axis in 3D TEE volumes. For more detailed definitions of the adopted metrics, we refer the reader to [26]. Metrics for the annulus are computed after the application of the reconstruction algorithm proposed in our previous work [33], which ensures consistent reconstruction of the whole annulus profile in case of under-segmented or noisy predictions. The tested hypothesis is that the predictions $\hat{y}_i$ yielded by the $S$ model are at least accurate as the predictions $\hat{y}_i$ by any $T_j$ or by $T$.

Second, effectively improving generalization to unlabeled data and providing reliable and temporally consistent frame-by-frame MV segmentations over the cardiac cycle. To this aim, each individual Teacher model $T_j$, the ensemble Teacher $T$, and the Student $S$ are evaluated separately on $D^u$ by comparing the respective predictions $\hat{y}_i$ vs. the GT segmentations $y_i$ for each $x_i \in D^u$ (Fig. 3). The metrics described in the previous paragraph are computed. The tested hypothesis is that the $S$ model generalizes on frames that are not ES or ED better than any $T_j$ and than $T$, thanks to the training on a larger and more heterogeneous set that also includes $\tilde{D}_u$. Additionally, the overall temporal consistency ($TCS$) of the predicted segmentation is evaluated across each frame in a cardiac cycle of the available 4D TEE recordings in the test set. Similarly to [24], temporal consistency is evaluated by measuring a temporal consistency score, $TCS_t$, for MV pre-

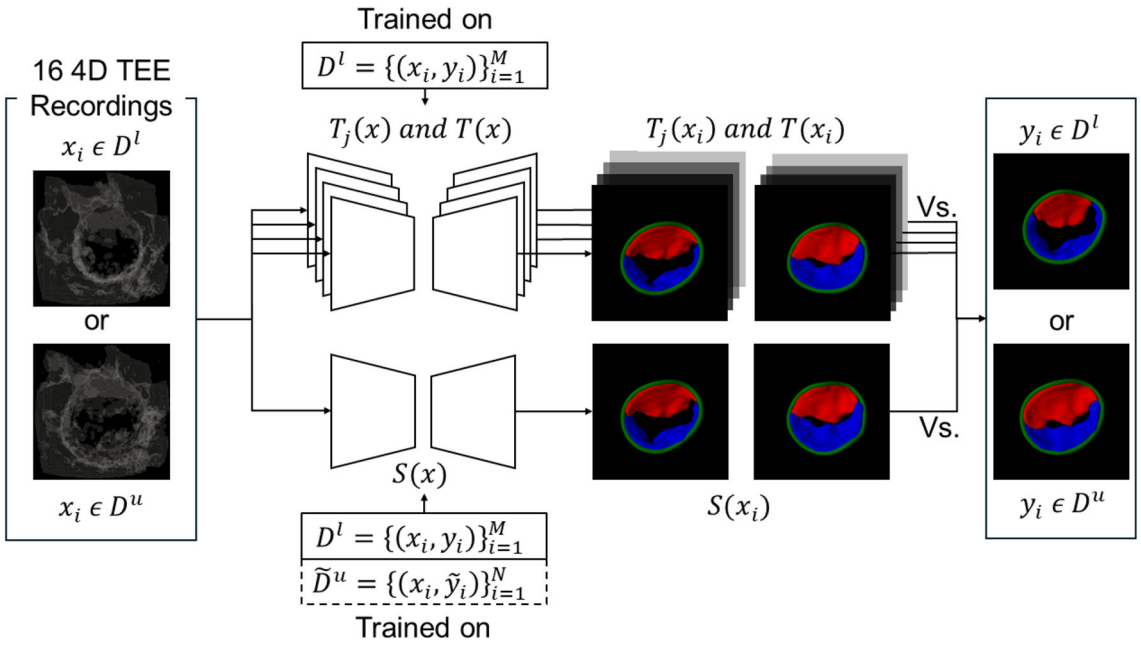


**Fig. 3** Schematic representation of the evaluation carried out on the 16 4D TEE recordings belonging to the test set. Inference of the individual Teachers $T_j$ and Ensemble $T$ is performed separately on both $D^l$ and $D^u$, and compared against Student $S$. Note that Teachers are only trained on a limited labeled dataset $D^l$ whereas the Student $S$ employs combined $D^l$ and $\tilde{D}^u$, leveraging a larger number of time frames

diction $\hat{y}_t$ at each frame $t$ of the 4D TEE recording. $TCS_t$ is defined as:

$$TCS_t = \begin{cases} 1, & \text{if } \hat{y}_t \text{ is consistent} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $\hat{y}_t$ is considered consistent if the following two conditions are satisfied:

- The mitral annulus is segmented in at least 70% of the long-axis views identified in the 3D TEE volume.
- The relative change in volume $\Delta V$ of the predicted anterior and posterior leaflets between two consecutive frames is less than a specified threshold $\delta$. This second condition leverages the incompressible behavior of leaflet tissue [44].

The first condition derives from the observation that the annulus reconstruction algorithm fails if the predicted segmentation of the annulus is missing from more than 30% of the long-axis views in the 3D TEE volume. The second condition verifies the physical consistency of the predicted 3D MV model by identifying the unacceptable volume changes in mitral leaflets predictions through time. The threshold $\delta$ is estimated as the average relative volume variation $\Delta V$ across the 4D TEE recordings in the test set, leveraging GT annotations of the frames sampled for constituting $D^l$ and $D^u$. $\delta$ is found to be 0.19 for the anterior leaflet and 0.15 for the posterior leaflet. $TCS$ is the fraction of frames in a recording that pass the automatic consistency evaluation [24] and is computed for the dynamic MV predictions of Ensemble $T$ and Student $S$, as well as the individual Teacher $T_3$, which shares the same SegResNet architecture with $S$. The segmentation consistency through time is further assessed by measuring the average absolute deviation (AAD) from the mean of the predicted anterior and posterior leaflet volumes across the cardiac cycle. This is given by

$$AAD_{mean} = \frac{1}{n_t} \sum_{t=1}^{n_t} |V_t - \bar{V}| \tag{2}$$

where $n_t$ is the total number of frames in a cardiac cycle, $V_t$ is the leaflet volume estimation at frame $t$, and $\bar{V}$ is the mean volume estimation throughout the cardiac cycle. Lower variations in the volume estimation from the predicted segmentations consequently mean higher temporal consistency.

All metrics, including temporal consistency, are computed with and without the use of SDA to address the impact of this regularization technique on the performance of the $S$ model.

## 4 Results

### 4.1 Performance on labeled dataset

Table 1 presents the average performance across the ES and ED frames in the test set of $D^l$. Overall, the Residual UNet ($T_1$), the Multi-Decoder Residual UNet ($T_2$), and the Seg-

**Table 1** Average segmentation performance across the ES and ED frames in the test set of $D^l$

| Metric | Teacher | | | | Student | |
|---|---|---|---|---|---|---|
| | $T_1$ | $T_2$ | $T_3$ | $T$ | $S$ w/o SDA | $S$ w/ SDA |
| Dice score | $0.790 \pm 0.075$ | $0.796 \pm 0.067$ | $0.796 \pm 0.063$ | $\mathbf{0.806 \pm 0.085}$ | $0.799 \pm 0.051$ | $\mathbf{0.806 \pm 0.049}$ |
| ASD (mm) | $0.501 \pm 0.284$ | $\mathbf{0.423 \pm 0.137}$ | $0.446 \pm 0.168$ | $0.425 \pm 0.191$ | $0.449 \pm 0.135$ | $0.442 \pm 0.139$ |
| 95%HD (mm) | $2.610 \pm 3.226$ | $2.281 \pm 1.684$ | $2.228 \pm 1.667$ | $2.034 \pm 1.300$ | $2.098 \pm 1.341$ | $\mathbf{2.019 \pm 1.258}$ |
| Curve-to-curve (mm) | $1.825 \pm 0.728$ | $2.009 \pm 1.071$ | $1.763 \pm 0.746$ | $\mathbf{1.686 \pm 1.120}$ | $1.920 \pm 0.903$ | $1.865 \pm 0.728$ |
| In-plane (mm) | $1.898 \pm 0.775$ | $2.441 \pm 1.755$ | $1.912 \pm 1.073$ | $\mathbf{1.818 \pm 3.252}$ | $2.231 \pm 1.571$ | $2.109 \pm 1.073$ |

Metrics include Dice score, ASD, 95%HD for the leaflets, and curve-to-curve error and in-plane error for the annulus. The Student $S$ is evaluated with and without the application of SDA during model training. The best values for each metric across all models are highlighted in bold

ResNet ($T_3$), trained exclusively on $D^l$, achieve comparable segmentation quality for the leaflets, with Dice score values exceeding 0.79. Notably, the Multi-Decoder Residual UNet ensures lower ASD and 95%HD values ($0.423 \pm 0.137$ mm and $2.281 \pm 1.684$ mm, respectively). For the annulus, the Residual UNet and SegResNet achieve curve-to-curve and in-plane errors less than 2 mm. As expected, the Ensemble $T$ outperforms the individual architectures, achieving higher Dice score values ($0.806 \pm 0.085$), comparable ASD values ($0.425 \pm 0.191$ mm), and lower 95%HD values ($2.034 \pm 1.300$ mm) for the leaflets. The Ensemble $T$ also ensures lower curve-to-curve error ($1.686 \pm 1.120$ mm) and in-plane error ($1.818 \pm 3.252$ mm) when identifying the annulus. The Student $S$ shows comparable metrics to the Ensemble $T$ and the single architectures in segmenting the MV at ES and ED frames. The use of SDA on $D^u$ slightly improves the performance on $D^l$ for the leaflets, yielding Dice score values of $0.806 \pm 0.049$, ASD values of $0.442 \pm 0.139$ mm, and 95%HD values of $2.019 \pm 1.258$ mm, matching the performance of the Ensemble $T$. When labeling the annulus, although the use of SDA on $D^u$ results in some improvements, the Student $S$ performs worse than the Ensemble $T$, with curve-to-curve and in-plane errors of $1.865 \pm 0.728$ mm and $2.109 \pm 1.073$, respectively. Figure 4 shows that the Student $S$, the Ensemble $T$, and the individual Teacher $T_3$ all yield MV segmentations that comparably match GT annotations when segmenting 3D TEE images at ES and ED. The Student $S$ identifies more precisely the boundaries between anterior and posterior leaflets at ES.

## 4.2 Performance on unlabeled dataset

Table 2 reports the average performance across the three intermediate frames in the test set of $D^u$. Remarkably, the Multi-Decoder Residual UNet ($T_2$) shows degraded performance when evaluated on the unseen intermediate frames of $D^u$, with lower Dice score values and higher distance metrics for both the leaflets and the annulus as compared

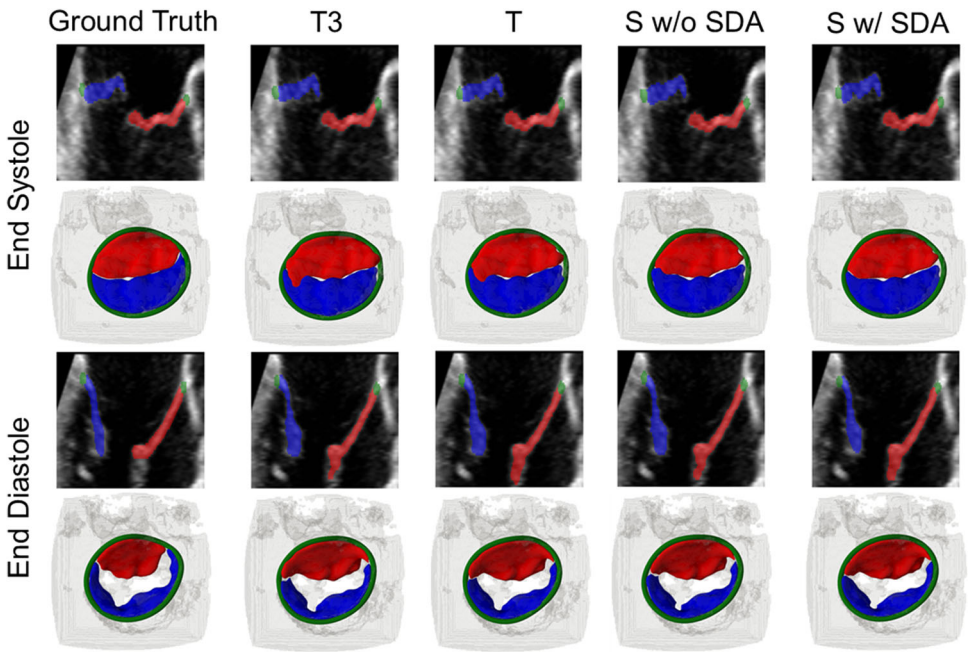


**Fig. 4** Segmentation examples from one 4D TEE recording in the test set. 2D views and 3D renderings of ES and ED frames from $D^l$ are shown along with the GT annotation and predictions from $T_3$, $T$, and $S$. Predictions for $S$ are shown both with and without the use of SDA. The annulus is labeled in green, the anterior leaflet in red, and the posterior leaflet in blue

**Table 2** Evaluation performance across the three intermediate frames in the test set of unlabeled dataset $D^u$

| Metric | Teacher | | | | Student | |
|---|---|---|---|---|---|---|
| | $T_1$ | $T_2$ | $T_3$ | $T$ | $S$ w/o SDA | $S$ w/ SDA |
| Dice score | $0.799 \pm 0.058$ | $0.732 \pm 0.133$ | $0.802 \pm 0.048$ | $0.809 \pm 0.049$ | $0.810 \pm 0.054$ | $\mathbf{0.815 \pm 0.040}$ |
| ASD (mm) | $0.413 \pm 0.143$ | $0.383 \pm 0.097$ | $0.406 \pm 0.156$ | $0.382 \pm 0.101$ | $0.469 \pm 0.548$ | $\mathbf{0.373 \pm 0.091}$ |
| 95%HD (mm) | $2.061 \pm 1.719$ | $5.168 \pm 6.153$ | $2.135 \pm 1.590$ | $1.995 \pm 1.682$ | $2.059 \pm 2.352$ | $\mathbf{1.719 \pm 1.043}$ |
| Curve-to-curve (mm) | $1.406 \pm 0.440$ | $3.593 \pm 3.677$ | $1.444 \pm 0.479$ | $\mathbf{1.357 \pm 0.432}$ | $1.751 \pm 1.601$ | $1.391 \pm 1.321$ |
| In-plane (mm) | $1.632 \pm 0.999$ | $5.389 \pm 7.237$ | $1.691 \pm 1.581$ | $\mathbf{1.420 \pm 0.645}$ | $2.995 \pm 4.325$ | $2.428 \pm 4.084$ |

For the Student $S$, performance metrics are reported with and without the application of SDA during model training. The best values for each metric across all models are highlighted in bold

to the performance achieved on $D^l$. In contrast, the Residual UNet ($T_1$) and the SegResNet ($T_3$) demonstrate higher generalization capability on unseen frames as compared to $T_2$, achieving Dice score values exceeding 0.79, ASD values lower than 0.5 mm, and 95%HD around 2 mm for the leaflets. For the annulus, $T_1$ and $T_3$ achieve curve-to-curve and in-plane errors lower than 1.7 mm. The Ensemble $T$ consistently outperforms the individual architectures and generalizes well to the unseen intermediate frames of $D^u$. It achieves Dice score values of $0.809 \pm 0.049$, ASD values of $0.382 \pm 0.101$ mm, and 95%HD values of $1.995 \pm 1.682$ mm for the leaflets, along with low distance values for the annulus. The performance metrics achieved by the Ensemble $T$ on $D^u$ underscore the consistency provided in the generation of the pseudo-labels for the training of the Student $S$. As compared to the Ensemble $T$, the Student $S$ provides comparable performance metrics in terms of Dice score ($0.810 \pm 0.054$) for the leaflets, but slightly lower values in terms of distances, particularly for the annulus. Notably, the Student $S$ benefits from the use of SDA on $D^u$ during training, consistently improving the Dice score values up to $0.815 \pm 0.040$ and lowering the ASD and 95%HD values to $0.372 \pm 0.091$ mm and $1.719 \pm 1.043$ mm, respectively, surpassing the Ensemble $T$. For the annulus, all the metrics improve with SDA, although curve-to-curve and in-plane errors remain worse than those obtained by the Ensemble $T$. Figure 5 shows an illustrative comparison where the Student $S$ outperforms the Ensemble $T$ and the individual Teacher $T_3$ in segmenting the MV in a generic diastolic frame pertaining to $D^u$. The use of SDA on $D^u$ further improves the segmentation, making it more consistent with the GT annotations. However, in systole the better performance of the Student $S$ and the improvement introduced by the additional use of SDA are less evident.

### 4.3 Frame-by-frame segmentation

Figure 6 shows the effect of the proposed self-training strategy to achieve MV 4D segmentation from 4D TEE

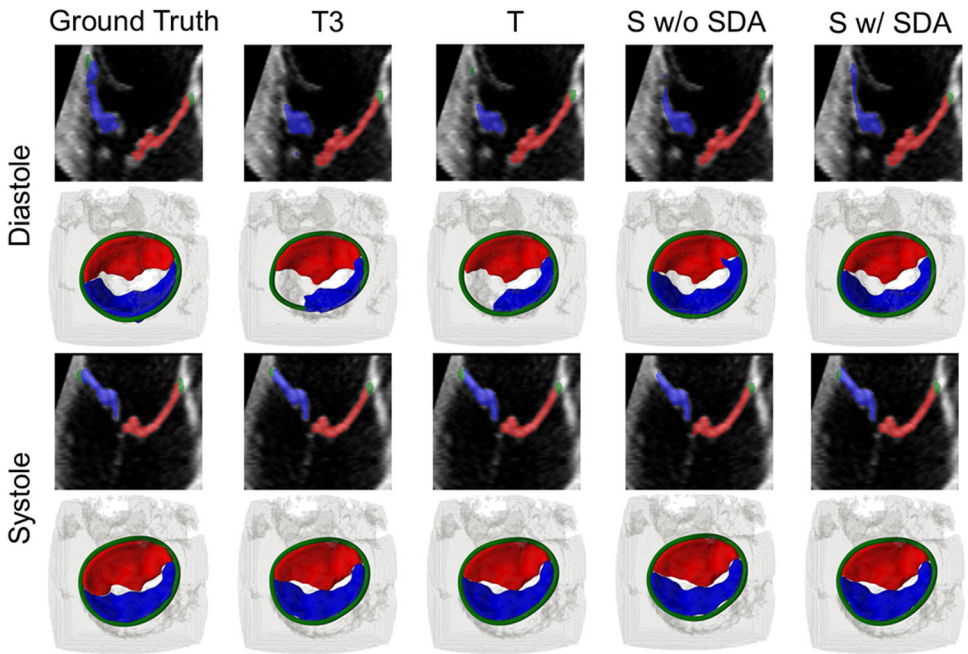


**Fig. 5** Segmentation examples from one 4D TEE recording in the test set. 2D views and 3D renderings of generic diastole and systole frames from $D^u$ are shown along with the GT annotation and predictions from $T_3$, $T$, and $S$. Predictions by the Student $S$ with and without the use of SDA are shown. The annulus is labeled in green, the anterior leaflet in red, and the posterior leaflet in blue

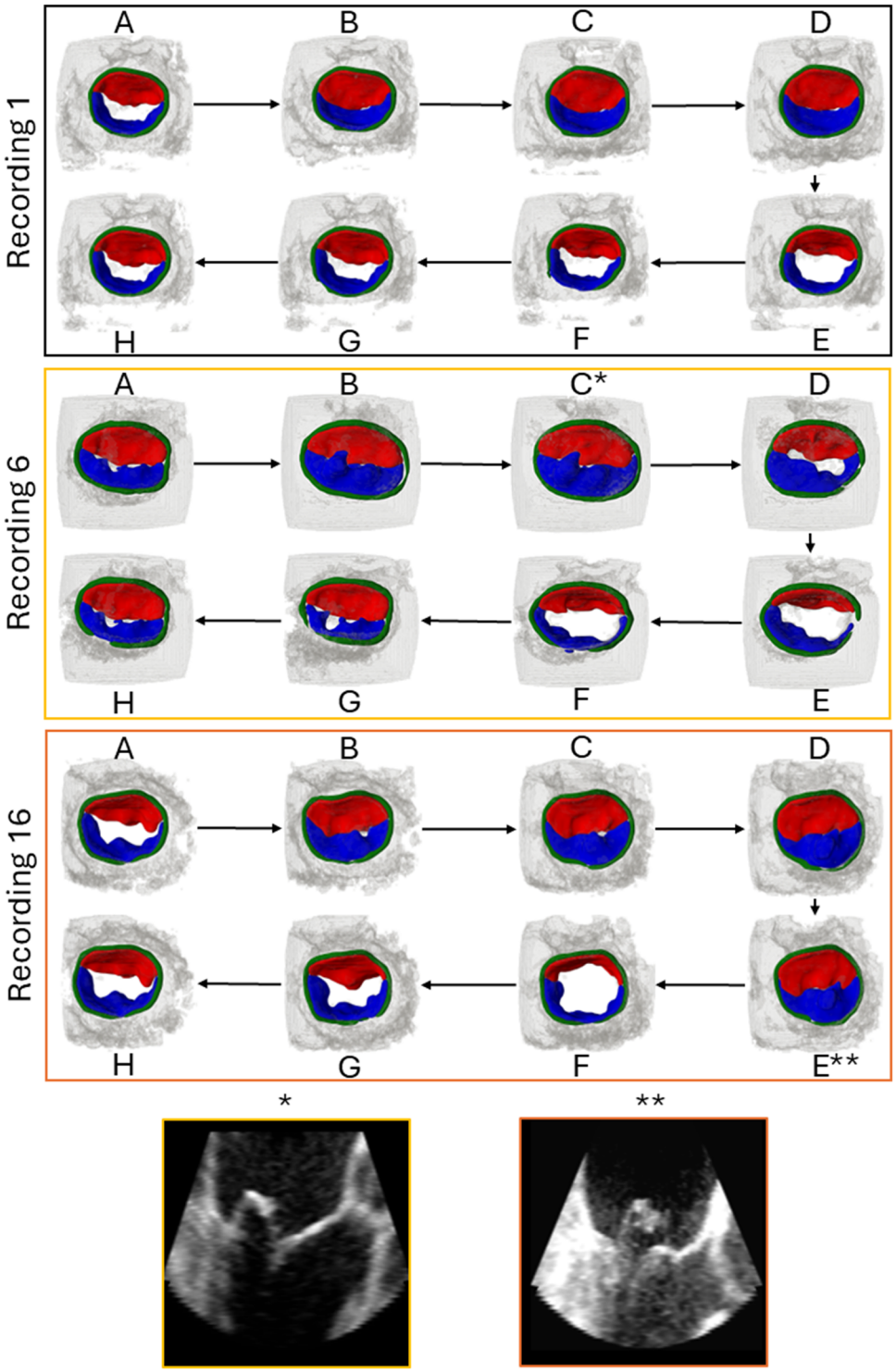


**Fig. 6** Automatic frame-by-frame segmentation from three randomly selected 4D TEE recordings. Each differently colored rectangle represents a distinct 4D TEE recording, with black arrows and letters indicating the chronological sequence. At the bottom, two long-axis views taken from systolic frame C of Recording 6 and frame E of Recording 16 highlight the prolapsing posterior leaflet profile correctly segmented by Student *S*. Color code: annulus is green, anterior leaflet is red, and posterior leaflet is blue

recordings. The Student *S* provides coherent estimation of the MV dynamics and morphology throughout the cardiac cycle in different 4D TEE recordings. Notably, the reconstruction of MV leaflets yielded by the Student *S* is accurate enough to allow for clearly assessing leaflet billowing or prolapse and coaptation gaps, when present, in the systolic frames, i.e., when the MV is closed (Fig. 6, Recording 6 and Recording 16). The predicted annulus presents discontinuities in some phases of MV dynamics, which can be resolved by applying the reconstruction algorithm proposed in [33]. Figure 7 compares TCS for the individual Teacher $T_3$, the Ensemble $T$, and the Student $S$. For clarity, only the individual Teacher

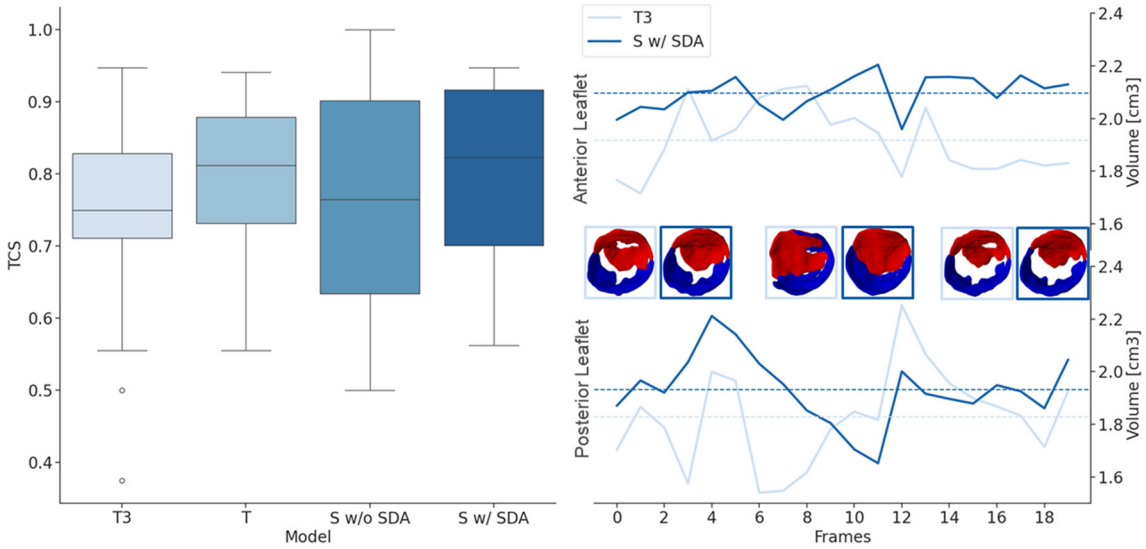


**Fig. 7** Boxplot of temporal consistency results for individual Teacher $T_3$, the Ensemble $T$, and the Student $S$ with and without the applications of SDA (left side). Illustrative example of predicted anterior and posterior leaflet volume predictions from a 4D TEE recording provided by $T_3$ and $S$ with SDA. Snapshots of MV segmentation from the frames are also shown (right side)

$T_3$ is considered in this analysis because it shares the same SegResNet architecture with the Student $S$ and provides better overall segmentations than the other individual Teachers on $D^l$ and $D^u$. The figure also exemplifies the variability in leaflets volume prediction for $T_3$ and $S$ (Fig. 7, right). The Student $S$ provides higher TCS values but with high variability across the test set (Fig. 7). The introduction of SDA on $D^u$ leads to overall better temporal consistency with median value of 0.823, exceeding both the individual Teacher $T_3$ (0.750, $p = 0.25$, $p > 0.05$) and the Ensemble $T$ (0.812, $p = 0.83$, $p > 0.05$). The predicted leaflet volumes in Fig. 7 (right panel), shows that the Student $S$ provides more temporally regular segmentations that are closer to the mean value than $T_3$, especially for the anterior leaflet, achieving a lower AAD of 0.054 $cm^3$ compared to 0.108 $cm^3$ of $T_3$. This is qualitatively observed in the snapshots representing the MV predictions from the two models at three consecutive time frames (Fig. 7, right). Table 3 reports AAD of the predicted leaflets volume across each frame of 4D TEE recordings in the test set. On average, the Student $S$ with SDA and the Ensemble $T$ achieve the lowest average absolute deviation for the anterior leaflets ($0.122 \pm 0.041$ $cm^3$ and $0.122 \pm 0.044$ $cm^3$, respectively). However, for the posterior leaflets, the use of SDA for training $S$ degrades leaflet consistency throughout the cardiac cycle, resulting in an average AAD of $0.192 \pm 0.081$ $cm^3$, matching the performance of the individual Teacher $T_3$.

Overall, the Student $S$ provides consistent temporal MV segmentations, requiring an average inference time of $0.182 \pm 0.038$ s per 4D TEE recording and $0.010 \pm 0.001$ s per single frame. Although, $S$ achieves comparable or slightly better MV temporal consistency than the Ensemble $T$, which requires $19.888 \pm 5.478$ s to infer 4D MV segmentation, and $1.169 \pm 0.272$ s per single frame.

## 5 Discussion

### 5.1 Novelty and added value of the proposed method

This study addresses a significant gap in automatic and standardized 4D TEE MV segmentation by introducing a novel application of a semi-supervised training strategy that leverages pseudo-labeling. The proposed approach reduces the reliance on manual annotations, while achieving high segmentation accuracy across the cardiac cycle, enhancing clinical applicability.

Although an exact comparison cannot be made, given the difference in the used dataset, the individual Teachers ($T_1$, $T_2$, $T_3$) report on $D^l$ results comparable to previous studies which focus only on MV segmentation at the ES frame [33], at the ED frame [21, 36] or at both frames [23] and report Dice score values ranging between 0.81 and 0.877, ASD values between 0.32 and 0. 92 mm, and 95% HD values between 1.86 and 1.99 mm. Similarly, the lowest curve-to-curve and in-plane errors obtained by $T_1$ and $T_3$ for the annulus are comparable to previous works reporting curve-to-curve and in-plane errors equal to 2.04 mm and 1.94 mm [25], or 1.96 mm and 1.82 mm [26]. However, the Ensemble $T$ consistently outperforms the individual Teachers $T_j$, ensuring higher Dice score (0.81) for the leaflets and lower curve-to-curve and in-plane errors (1.67 mm and 1.81 mm, respectively). This demonstrates the advantage of ensemble methods in providing more

**Table 3** Average absolute deviation of predicted anterior leaflet (AL) and posterior leaflet (PL) volume estimation over the cardiac cycle

| Average Absolute Deviation | $T_3$ | $T$ | $S$ w/o SDA | $S$ w/ SDA |
|---|---|---|---|---|
| AL Volume ($cm^3$) | $0.148 \pm 0.039$ | $0.122 \pm 0.044$ | $0.135 \pm 0.060$ | $\mathbf{0.122 \pm 0.041}$ |
| PL Volume ($cm^3$) | $0.192 \pm 0.081$ | $0.175 \pm 0.073$ | $\mathbf{0.165 \pm 0.093}$ | $0.192 \pm 0.088$ |

Mean and standard deviation are computed across each 4D TEE recording of the test set and reported for model $T_3$, the Ensemble $T$ and the Student model $S$ with and without the applications of SDA
The lowest average absolute deviations of predicted anterior leaflet and posterior leaflet volume estimation are highlighted in bold

accurate and robust predictions by averaging out biases [34]. The use of the pseudo-labels, produced by the Ensemble $T$ for $D^u$, in training the Student $S$ does not negatively impact the performance of the Student $S$ on $D^l$; instead, it ensures comparable performance with respect to the Ensemble $T$, especially with SDA applied during training. This indicates that the Ensemble $T$ generates reliable pseudo-labels, and SDA mitigates potential errors in pseudo-labeling. Quantitative evaluation of the negative impact of label noise in deep neural network training is achievable through various techniques [47, 48]. However, these methods require a subset of data free of label noise or at least knowledge of the noise level in labels, information that is often prohibitively costly to obtain. Stochastic co-teaching has recently been proposed to address this limitation. It involves sequentially training two models with identical architectures on the same data but with different instance selections. In this framework, each model selects training instances for the other based on GT posterior probabilities and a stochastically chosen threshold [49]. This method eliminates the need for assumptions regarding the expected level of label noise. Nonetheless, we believe that an in-depth analysis of label noise impact would deviate from the main scope of this work, especially given the favorable results on the labelled dataset ensured through pseudo labelling using ensemble modelling. Previous studies [50, 51] have applied an ensemble method in self-training strategies to mitigate the risk of incorrect pseudo-label, which may cause severe error propagation referred to as confirmation bias. In these works, each member of the ensemble consisted of the same network, but was trained with different stochastic input augmentation. Here, we chose to exploit three different CNN architectures for ensemble modeling. By combining their strengths and compensating for their weakness we effectively address error propagation. The proposed self-training strategy is then effective in augmenting the training dataset for the Student $S$, enhancing segmentation performance.

The evaluation on $D^u$ highlights the generalization ability over multiple frames of the cardiac cycle, i.e., the capability to correctly segment MV substructures in open, closed, or intermediate configurations of the MV. The Multi-Decoder Residual UNet ($T_2$) exhibits degraded performance on unseen time frames, likely due to the higher complexity of its architecture that makes it more prone to overfitting. In contrast, $T_1$ and $T_3$ demonstrate good generalization abilities, overall maintaining the same performance showed on $D^l$. As expected, the Ensemble $T$ again outperforms the individual Teachers $T_j$ for both mitral leaflets and annulus segmentation in terms of Dice score, ASD, and 95%HD values. Notably, the Student $S$ overall equals the segmentation performance of the Ensemble $T$ but improves the mitral leaflets Dice score values, suggesting higher generalization capabilities and more reliable MV segmentation than $T$. SDA further improves the results for leaflets segmentation, guaranteeing the lowest ASD and 95%HD values, and the best Dice score. The results achieved on $D^u$ demonstrate the effectiveness of the proposed self-training strategy in improving generalization performance as compared to the Teachers. We believe that this improvement is due not only to the additional training examples provided for the Student $S$ but also to the choice of training a randomly initialized model on a combined labeled and pseudo-labeled dataset from the start. This approach helps the Student $S$ learning more generalized patterns by introducing variability at the beginning of the training process. Randomly initialized weights have been found to be effective in improving generalization capability and mitigating the performance lag typically experienced in two-phase learning regimes using pre-trained models [52, 53]. Therefore, this choice acts as a regularization technique which, together with the use of SDA [32], contributes to learning robust features from pseudo-labeled data.

The proposed strategy achieves unprecedented automatic 4D MV segmentation, with the Student $S$ providing consistent segmentations across the cardiac cycle, accurately detecting anatomical features of regurgitant MV, such as prolapses (Fig. 6). Indeed, the Student $S$ provides better $TCS$, when SDA is applied, in 4D MV segmentation as compared to the individual Teacher $T_3$, which shares the same SegResNet architecture with Student $S$, and to the ensemble $T$. Although these differences are not statistically significant ($p > 0.05$), probably due to the limited sample size in the test set (16 4D TEE recordings), the Student $S$ exhibits fewer fluctuations in predicted leaflet volumes across frames (Fig. 7), confirmed by a low mean AAD for anterior leaflet volume and for posterior leaflet volume. These results confirm the effectiveness of the proposed method in improving the generalization performance. The Student $S$ is better equipped to capture the rapid changes in MV configurations though each frame of a 4D TEE recording regardless of the temporal reso-

lution considered. This leads to a more reliable MV dynamic segmentation. However, the application of SDA contributes to reducing the AAD for the anterior leaflet but not for the posterior leaflet. Of note, the MV posterior leaflet exhibits particularly high shape variability in 4D TEE recordings which is challenging to address in automatic segmentation methods. Indeed, the posterior leaflet often opens toward the direction of the ultrasound waves, which determines loss of information due to signal dropout [46]. Consequently, this aspect may influence the analysis of temporal consistency, which is based on the assumption of constant leaflet volume during the cardiac cycle, because leaflet tissue is incompressible [44], and any permissible variation in volume of the segmentation mask must be associated with limitations of echocardiography, such as signal dropout.

To the best of our knowledge, this is the first work which presents and evaluates an automatic method for 4D MV with focus on regurgitant MVs. Unlike a previous work that only qualitatively demonstrated 4D MV segmentation [23], our work extensively and quantitatively evaluates the proposed strategy, comparing predicted results with expert manual annotations and introducing criteria for temporal consistency evaluation. Moreover, the multi-class segmentation task enriches the value of the proposed strategy by providing separate segmentations of the mitral annulus, anterior, and posterior leaflets. Finally, as compared to the Ensemble $T$ the Student $S$ is not only comparably reliable, but also almost 100 times faster, on average, in segmenting the 4D TEE data, as detailed in Sect. 4.3

### 5.2 Clinically relevant fallouts

The accuracy, automation, and time-efficiency of the proposed Student $S$, as well as its capability to provide frame-by-frame segmentations of MV substructures, make it a potentially useful tool to support diagnosis and procedural planning.

Indeed, two main types of MR exist, which are characterized by completely different pathophysiological mechanisms and require different management [55]: organic, which is related to primary alterations of MV substructures, and functional, which is secondary to the adverse remodeling of the left ventricle. The most common form of organic MR is degenerative, which encompasses a broad range of alterations in MV tissues.

The patients selected in the present study were all affected by degenerative MR. When managing this type of patients, the detailed quantification of MV morphology and dynamics over the cardiac cycle is key to correctly classifying MR etiology within the spectrum of degenerative conditions, and hence to selecting the most suitable MV repair procedure [54]. The approach herein proposed would allow for such a detailed quantification based on 4D TEE: it allows not only identify leaflet billowing and regurgitation gaps at ES, when the MV would be perfectly sealed if healthy, but also to assess MV dynamics thanks to reliable frame-by-frame segmentation. Also, the full automation of the segmentation and reconstruction process would make the quantification operator-independent and fully repeatable, while decreasing the workload of sonographers and improving clinical workflow efficiency.

The time required by the Student $S$ to process a single cardiac cycle in a 4D TEE recording make it suitable for intra-procedural use. In particular, it could be useful in transcatheter MV repair, which is performed in beating hearts and requires the continuous monitoring, through 4D TEE, of MV position and motion in the intracardiac space [56].

Additionally, the automatic reconstruction analysis of MV dynamics from 4D TEE could be used to improve numerical models for the quantification of MV biomechanics, which are proposed in the literature as a means to better tailor surgical or transcatheter treatments to the patient [57, 58]. The data yielded by the Student $S$ could be used to feed those models, by providing the unloaded MV geometry at ED and annular motion to be imposed as boundary condition or to validate them, by providing the GT time-dependent configuration of the MV.

### 5.3 Limitations and future works

The choice of Teacher architectures affects the generation of pseudo-labels, requiring models with good generalization abilities. Exploring other architectures, such as Attention UNet, which proved robust against image artifacts [24], could improve pseudo-label accuracy and robustness against signal dropout-related issues. Additionally, the proposed method is validated on a relatively small, homogeneous dataset of MR patients. Future work should include larger, more heterogeneous datasets, e.g., including MR with other etiologies, to confirm robustness. Furthermore, incorporating MV temporal consistency as a regularization technique in the training process could enhance dynamic segmentation. Finally, integrating the segmentation framework with real-time TEE imaging systems could facilitate its use in intraoperative settings, providing surgeons with real-time feedback during MV examinations or repair procedures.

## 6 Conclusion

We proposed a semi-supervised segmentation strategy that significantly advances 4D TEE MV segmentation by reducing the need for extensive manual annotations while achieving high segmentation accuracy and temporal consistency. This work formulates a self-training strategy with pseudo-labeling, based on the Teacher-Student framework, to address the lack of training data for dynamic MV segmentation in 4D

TEE. The generation of pseudo-labels using an ensemble of three different CNN architectures and the application of a peculiar SDA on the pseudo-labeled dataset were effective in increasing the generalization performance for temporal segmentation of the Student model. Our approach has strong potential for clinical applications.

**Funding** Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement. This work was supported by the European Union's Horizon 2020 research and innovation program, under the project ARTERY, grant agreement No. 101017140.

## Declarations

**Conflict of interest** The authors declare no competing interests.



## References

1. El Sabbagh A, Reddy Y, Nishimura R (2018) Mitral valve regurgitation in the contemporary era: insights into diagnosis, management, and future directions. JACC Cardiovasc Imaging 11:628–643
2. Nkomo V, Gardin J, Skelton T, Gottdiener J, Scott C, Enriquez-Sarano M (2006) Burden of valvular heart diseases: a population-based study. Lancet 368:1005–1011
3. Otto C, Nishimura R, Bonow R, Carabello B, Erwin J III, Gentile F, Jneid H, Krieger E, Mack M, McLeod C et al (2021) 2020 ACC/AHA guideline for the management of patients with valvular heart disease: executive summary: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. J Am Coll Cardiol 77:450–500
4. Clavel M, Mantovani F, Malouf J, Michelena H, Vatury O, Jain M, Mankad S, Suri R, Enriquez-Sarano M (2015) Dynamic phenotypes of degenerative myxomatous mitral valve disease: quantitative 3-dimensional echocardiographic study. Circ Cardiovasc Imaging 8:e002989
5. Wijngaarden S, Kamperidis V, Regeer M, Palmen M, Schalij M, Klautz R, Bax J, Ajmone Marsan N, Delgado V (2018) Three-dimensional assessment of mitral valve annulus dynamics and impact on quantification of mitral regurgitation. Eur Heart J Cardiovasc Imaging 19:176–184
6. Suri R, Vanoverschelde J, Grigioni F, Schaff H, Tribouilloy C, Avierinos J, Barbieri A, Pasquet A, Huebner M, Rusinaru D et al (2013) Association between early surgical intervention vs watchful waiting and outcomes for mitral regurgitation due to flail mitral valve leaflets. Jama. 310:609–616
7. Zamorano J, Badano L, Bruce C, Chan K, Gonçalves A, Hahn R, Keane M, La Canna G, Monaghan M, Nihoyannopoulos P et al (2011) EAE/ASE recommendations for the use of echocardiography in new transcatheter interventions for valvular heart disease. Eur Heart J. 32:2189–2214
8. Sorrentino S, Berardini A, Statuto G, Angeletti A, Massaro G, Capobianco C, Piemontese G, Spadotto A, Toniolo S, Caponetti A et al (2021) Percutaneous mitral valve repair with the mitraclip system in the current clinical practice. Hearts 2:74–86
9. Ginty O, Moore J, Eskandari M, Carnahan P, Lasso A, Jolley M, Monaghan M, Peters T (2019) Dynamic, patient-specific mitral valve modelling for planning transcatheter repairs. Int J Comput Assist Radiol Surg 14:1227–1235
10. Rim Y, McPherson D, Chandran K, Kim H (2013) The effect of patient-specific annular motion on dynamic simulation of mitral valve function. J Biomech 46:1104–1112
11. Votta E, Caiani E, Veronesi F, Soncini M, Montevecchi F, Redaelli A (2008) Mitral valve finite-element modelling from ultrasound data: a pilot study for a new approach to understand mitral function and clinical scenarios. Philo Trans R Soc A Math Phys Eng Sci 366:3411–3434
12. Hien M, Großgasteiger M, Weymann A, Rauch H, Rosendal C (2014) Reproducibility in echocardiographic two-and three-dimensional mitral valve assessment. Echocardiography 31:311–317
13. Thomas N, Unsworth B, Ferenczi E, Davies J, Mayet J, Francis D (2008) Intraobserver variability in grading severity of repeated identical cases of mitral regurgitation. Am Heart J 156:1089–1094
14. Ionasec R, Voigt I, Georgescu B, Wang Y, Houle H, Vega-Higuera F, Navab N, Comaniciu D (2010) Patient-specific modeling and quantification of the aortic and mitral valves from 4-D cardiac CT and TEE. IEEE Trans Med Imaging 29:1636–1651
15. Burlina P, Sprouse C, DeMenthon D, Jorstad A, Juang R, Contijoch F, Abraham T, Yuh D, McVeigh E (2010) Patient-specific modeling and analysis of the mitral valve using 3D-TEE. Information Processing In Computer-Assisted Interventions: First International Conference, IPCAI 2010, Geneva, Switzerland, June 23, 2010. Proceedings 1. pp. 135-146
16. Schneider R, Burke W, Marx G, Nido P, Howe R (2011) Modeling mitral valve leaflets from three-dimensional ultrasound. Functional Imaging And Modeling Of The Heart: 6th International Conference, FIMH 2011, New York City, NY, USA, May 25-27, 2011. Proceedings 6. pp. 215-222
17. Pouch AM, Yushkevich PA, Jackson BM, Jassar AS, Vergnat M, Gorman JH, Gorman RC, Sehgal CM (2012) Development of a semi-automated method for mitral valve modeling with medial axis representation using 3D ultrasound. Med Phys. 39:933–950
18. Pouch A, Wang H, Takabe M, Jackson B, Gorman J III, Gorman R, Yushkevich P, Sehgal C (2014) Fully automatic segmentation of the mitral leaflets in 3D transesophageal echocardiographic images using multi-atlas joint label fusion and deformable medial modeling. Med Image Anal 18:118–129
19. Pouch A, Aly A, Lai E, Yushkevich N, Stoffers R, Gorman J, Cheung A, Gorman J, Gorman R, Yushkevich P (2017) Spatiotemporal segmentation and modeling of the mitral valve in real-time 3D echocardiographic images. Medical Image Computing And Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20. pp. 746-754
20. Costa E, Martins N, Sultan M, Veiga D, Ferreira M, Mattos S, Coimbra M (2019) Mitral valve leaflets segmentation in echocardiography using convolutional neural networks. 2019 IEEE 6th Portuguese Meeting On Bioengineering (ENBENG). pp 1–4
21. Carnahan P, Moore J, Bainbridge D, Eskandari M, Chen E, Peters T (2021) DeepMitral: Fully automatic 3D echocardiography segmentation for patient specific mitral valve modelling. Medical Image Computing And Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part V 24. pp. 459-468

22. Aly A, Khandelwal P, Aly A, Kawashima T, Mori K, Saito Y, Hung J, Gorman J III, Pouch A, Gorman R et al (2022) Fully automated 3d segmentation and diffeomorphic medial modeling of the left ventricle mitral valve complex in ischemic mitral regurgitation. Med Image Anal 80:102513
23. Chen J, Li H, He G, Yao F, Lai L, Yao J, Xie L (2023) Automatic 3D mitral valve leaflet segmentation and validation of quantitative measurement. Biomed Sig Process Control 79:104166
24. Wifstad S, Kildahl H, Grenne B, Holte E, Hauge S, Sæbø S, Mekonnen D, Nega B, Haaverstad R, Estensen M et al (2024) Mitral valve segmentation and tracking from transthoracic echocardiography using deep learning. Ultrasound Med Biol
25. Andreassen B, Veronesi F, Gerard O, Solberg A, Samset E (2019) Mitral annulus segmentation using deep learning in 3-D transesophageal echocardiography. IEEE J Biomed Health Infor 24:994–1003
26. Andreassen B, Völgyes D, Samset E, Solberg A (2022) Mitral annulus segmentation and anatomical orientation detection in TEE images using periodic 3D CNN. IEEE Access. 10:51472–51486
27. Taskén A, Berg E, Grenne B, Holte E, Dalen H, Stølen S, Lindseth F, Aakhus S, Kiss G (2023) Automated estimation of mitral annular plane systolic excursion by artificial intelligence from 3D ultrasound recordings. Artif Intell Med 144:102646
28. Wu P, Huang Q, Yi J, Qu H, Ye M, Axel L, Metaxas D (2020) Cardiac MR image sequence segmentation with temporal motion encoding. Computer Vision-ECCV 2020 Workshops: Glasgow, UK, August 23-28, 2020, Proceedings, Part I 16. pp 298-309
29. Deng Y, Cai P, Zhang L, Cao X, Chen Y, Jiang S, Zhuang Z, Wang B (2022) Myocardial strain analysis of echocardiography based on deep learning. Front Cardiovasc Med 9:1067760
30. Lu J, Millioz F, Varray F, Porée J, Provost J, Bernard O, Garcia D, Friboulet D (2023) Ultrafast cardiac imaging using deep learning for speckle-tracking echocardiography. IEEE Transactions On Ultrasonics, Ferroelectrics, And Frequency Control
31. Ouali Y, Hudelot C, Tami M (2020) An overview of deep semi-supervised learning. arXiv:2006.05278
32. Yang L, Zhuo W, Qi L, Shi Y, Gao Y (2022) St++: make self-training work better for semi-supervised semantic segmentation. Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition. pp. 4268-4277
33. Munafò R, Saitta S, Ingallina G, Denti P, Maisano F, Agricola E, Redaelli A, Votta E (2024) A deep learning-based fully automated pipeline for regurgitant mitral valve anatomy analysis from 3D Echocardiography. IEEE Access
34. Li H, Jiang G, Zhang J, Wang R, Wang Z, Zheng W, Menze B (2018) Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. Neuroimage 183:650–665
35. DeVries T, Taylor G (2017) Improved regularization of convolutional neural networks with cutout. arXiv:1708.04552
36. Carnahan P, Bharucha A, Eskandari M, Chen E, Peters T (2023) Segmentation of the mitral valve from 3D transesophageal echocardiography. International Conference On Medical Image Computing And Computer Assisted Intervention (MICCAI) 2023 (MICCAI 2023). https://doi.org/10.5281/zenodo.7844870
37. Kerfoot E, Clough J, Oksuz I, Lee J, King A, Schnabel J (2019) Left-ventricle quantification using residual U-net. Statistical atlases and computational models of the heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held In Conjunction With MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 9. pp 371-380
38. Siddique M, Yang D, He Y, Xu D, Myronenko A (2022) Automated ischemic stroke lesion segmentation from 3D MRI. arXiv:2209.09546
39. Ginsburg B, Castonguay P, Hrinchuk O, Kuchaiev O, Lavrukhin V, Leary R, Li J, Nguyen H, Zhang Y, Cohen J (2019) Stochastic gradient methods with layer-wise adaptive moments for training of deep networks. arXiv:1905.11286
40. Yeung M, Sala E, Schönlieb C, Rundo L (2022) Unified focal loss: generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Comput Med Imaging Graph 95:102026
41. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al (2019) PyTorch: an imperative style, high-performance deep learning library. Adv Neural Infor Process Syst 32
42. Cardoso M, Li W, Brown R, Ma N, Kerfoot E, Wang Y, Murrey B, Myronenko A, Zhao C, Yang D, et al (2022) MONAI: an open-source framework for deep learning in healthcare. arXiv:2211.02701
43. Kikinis R, Pieper S, Vosburgh K (2013) 3D Slicer: a platform for subject-specific image analysis, visualization, and clinical support. Intraoperative Imaging And Image-guided Therapy. pp 277–289
44. May-Newman K, Yin F (1998) A Constitutive law for mitral valve tissue. J Biomech Eng 120:38–47. https://doi.org/10.1115/1.2834305
45. Sohn K, Berthelot D, Carlini N, Zhang Z, Zhang H, Raffel C, Cubuk E, Kurakin A, Li C (2020) FixMatch: simplifying semi-supervised learning with consistency and confidence. Adv Neural Inf Process Syst 33:596–608
46. Xia W, Moore J, Chen E, Xu Y, Ginty O, Bainbridge D, Peters T (2018) Signal dropout correction-based ultrasound segmentation for diastolic mitral valve modeling. J Med Imaging 5:021214–021214
47. Song H, Kim M, Park D, Shin Y, Lee J (2022) Learning from noisy labels with deep neural networks: a survey. IEEE Trans Neural Netw Learn Syst 34:8135–8153
48. Karimi D, Dou H, Warfield S, Gholipour A (2020) Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med Image Anal 65:101759
49. Vos B, Jansen G, Išgum I (2023) Stochastic co-teaching for training neural networks with unknown levels of label noise. Sci Rep 13:16875
50. Ahmed W, Morerio P, Murino V (2021) Adaptive pseudo-label refinement by negative ensemble learning for source-free unsupervised domain adaptation. arXiv:2103.15973
51. Wang K, Zhang C, Geng Y, Ma H (2024) Evidential Pseudo-Label Ensemble for semi-supervised classification. Pattern Recogn Lett. 177:135–141. https://www.sciencedirect.com/science/article/pii/S0167865523003379
52. Ash J, Adams R (2020) On warm-starting neural network training. Adv Neural Inf Process Syst 33:3884–3894
53. Berariu T, Czarnecki W, De S, Bornschein J, Smith S, Pascanu R, Clopath C (2021) A study on the plasticity of neural networks. arXiv:2106.00042
54. Sturla F, Onorati F, Puppini G, Pappalardo O, Selmi M, Votta E, Faggian G, Redaelli A (2017) Dynamic and quantitative evaluation of degenerative mitral valve disease: a dedicated framework based on cardiac magnetic resonance imaging. J Thorac Dis 9:S225–S238
55. De Bonis M, Maisano F, Canna G, Alfieri O (2012) Treatment and management of mitral regurgitation. Nat Rev Cardiol 9:133–146
56. Ramchand J, Harb S, Krishnaswamy S, Jaber W, Miyasaka R (2020) Echocardiographic guidance of transcatheter mitral valve edge-to-edge repair. Structural Heart. 4:397–412
57. Sturla F, Vismara R, Jaworek M, Votta E, Romitelli P, Pappalardo O, Lucherini F, Antona C, Fiore G, Redaelli A (2017) In vitro and in silico approaches to quantify the effects of the Mitraclip® system on mitral valve function. J Biomech 50:83–92
58. Sacks M, Drach A, Lee C, Khalighi A, Rego B, Zhang W, Ayoub S, Yoganathan A, Gorman R, Gorman J III (2019) On the simulation

of mitral valve function in health, disease, and treatment. J Biomech Eng 141:070804



**Riccardo Munafò** is PhD candidate in Bioengineering at Politecnico di Milano. His research activity focuses on AI-based methods for real-time segmentation of cardiac imaging and intracardiac catheter tracking.

**Simone Saitta** is Assistant Professor at University of Amsterdam, developing his research in the field of machine learning for medical image analysis and computational biomechanics.

**Davide Tondi** is PhD candidate in Bioengineering at Politecnico di Milano. His research activity focuses on finite element modelling for the clinical evaluation of valve pathologies and support of atrioventricular valve repair.

**Giacomo Ingallina** is Consultant Cardiologist in the Cardiovascular Imaging Unit at San Raffaele Hospital, with research interest devoted to the comprehensive utilization of echocardiography in valvular heart diseases.

**Paolo Denti** is responsible for transcatheter therapy as a hybrid cardiovascular surgeon at San Raffaele Hospital, with research interests in the field of percutaneous approach to structural heart disease.

**Francesco Maisano** is Full Professor of Cardiac Surgery at Vita-Salute University. His research activities primarily focus on the percutaneous approach to structural heart disease.

**Eustachio Agricola** is Associate Professor of Cardiology of Vita-Salute University. His research activity is mainly focused on the field of Cardiovascular Imaging, specializing in echocardiography.

**Emiliano Votta** is Associate Professor in Bioengineering at Politecnico di Milano. His main research activity focuses on the quantitative analysis and modeling of cardiovascular biological structures based on clinical imaging.