Computer Methods and Programs in Biomedicine 272 (2025) 109083

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

MitraClip device automated localization in 3D transesophageal echocardiography via deep learning☆

Riccardo Munafò [a], Simone Saitta [a,b,c], Luca Vicentini [d], Davide Tondi [a], Veronica Ruozzi [a,*], Francesco Sturla [e,a], Giacomo Ingallina [f], Andrea Guidotti [d], Eustachio Agricola [f,g], Emiliano Votta [a]

[a] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy
[b] Department of Biomedical Engineering and Physics, Amsterdam UMC, Amsterdam, The Netherlands
[c] Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
[d] Simulands, Zurich, Switzerland
[e] 3D and Computer Simulation Laboratory, IRCCS Policlinico San Donato, Milan, Italy
[f] Unit of Cardiovascular Imaging, IRCCS San Raffaele Hospital, Milan, Italy
[g] Vita-Salute San Raffaele University, Milan, Italy

ARTICLE INFO

Keywords:
Three-dimensional transesophageal echocardiography
Mitral valve
Mitral regurgitation
Percutaneous interventions
Transcatheter edge-to-edge repair
Automatic segmentation
Convolutional neural network

ABSTRACT

Background and Objective: The MitraClip is the most widely used percutaneous treatment for mitral regurgitation, typically performed under the real-time guidance of 3D transesophageal echocardiography (TEE). However, artifacts and low image contrast in echocardiography hinder accurate clip visualization. This study presents a proof-of-concept of an automated pipeline for clip detection from 3D TEE images acquired in a controlled in vitro simulation environment.

Methods: An Attention UNet was employed to segment the device, while a DenseNet classifier predicted its configuration among ten possible states, ranging from fully closed to fully open. Based on the predicted configuration, a template model derived from computer-aided design (CAD) was automatically registered to refine the segmentation and enable quantitative characterization of the device. The pipeline was trained and validated on 196 3D TEE images acquired using a heart simulator, with ground-truth annotations refined through CAD-based templates.

Results: The Attention UNet achieved an average surface distance of 0.76 mm and a 95% Hausdorff distance of 2.44 mm for segmentation, while the DenseNet achieved an average weighted F1-score of 0.80 for classification. Post-refinement, segmentation accuracy improved, with average surface distance and 95% Hausdorff distance reduced to 0.69 mm and 1.83 mm, respectively.

Conclusion: This pipeline enhanced clip visualization, providing fast and accurate detection with quantitative feedback, potentially improving procedural efficiency and reducing adverse outcomes.

1. Introduction

Transcatheter edge-to-edge repair (TEER) is the most widespread percutaneous treatment for mitral regurgitation (MR) [1]. It offers a safe and effective alternative for patients with contraindications for surgery or those at high operative risk [2,3]. The MitraClip (Abbott Laboratories, California, USA) is a catheter-based technology for mitral valve (MV) TEER, designed to treat MR by clipping together the mitral leaflets. The procedure involves accessing the left atrium (LA) through the interatrial septum with a steerable sheath. A delivery catheter is then advanced in the sheath and steered through the LA to reach the MV region where lack of coaptation is observed. Upon reaching this target region, a clip, located on the tip of the delivery catheter, is actuated to grasp the MV leaflets and it is finally deployed. To achieve optimal grasping, the clip should be oriented perpendicularly to the MV, with its arms locally orthogonal to the coaptation line to be restored.

These steps, from transseptal puncture through the steering of the delivery catheter toward MV and to clip positioning, are performed

☆ This work was supported by the European Union's Horizon 2020 research and innovation program, under the project ARTERY, grant agreement No. 101017140.

* Corresponding author.
E-mail address: veronica.ruozzi@polimi.it (V. Ruozzi).

https://doi.org/10.1016/j.cmpb.2025.109083
Received 2 February 2025; Received in revised form 3 September 2025; Accepted 22 September 2025
Available online 27 September 2025

under the guidance of 2D and 3D transesophageal echocardiography (TEE) [4]. 3D TEE provides an *en face view* of the MV and of the approaching clip, allowing clinicians to determine when the clip is adequately positioned and oriented with respect to the target region of the MV leaflets [4]. 3D TEE images can be non-trivial to interpret, since they are displayed on 2D monitors, with a loss of depth perception. Thus, clinicians often rely on 2D views for clip positioning, which requires adjusting the TEE transducer to align the clip in 2D TEE views or extracting specific slices from the 3D volume for better visualization on the fly [5]. However, both actions are time-consuming and operator-dependent, and require specific expertise. Moreover, the presence of the catheter and of the clip generates artifacts that hamper the assessment of clip configuration, in particular when it is close to native tissues. To address these challenges, an automated, real-time clip detection system for 3D TEE could provide significant benefits. Such a system would enable the automatic extraction of relevant 2D views encompassing the device, and, when integrated with automated MV analysis [6], facilitate precise quantification of the spatial relationship between the clip and its target region. This would enhance procedural guidance by providing standardized, operator-independent visualization.

1.1. Related works

Despite the clinical importance of intraoperative guidance, no prior study has tackled the automated segmentation of TEE for characterizing the position and configuration of clip devices for TEER. Published studies have primarily focused on catheter localization in TEE images, leveraging detection or segmentation techniques [7]. Detection methods aim to identify the general location or shape of the catheter by using representations such as bounding box (a rectangular region enclosing the catheter), axis (a central line running along the catheter's length), or skeleton (a simplified line-based representation capturing the catheter's structure). These methods typically rely on carefully designed filters or instrument templates to extract catheter-related information from the image [7]. On the other hand, segmentation involves classifying each voxel associated with the catheter to reconstruct the latter within the image. Convolutional neural networks (CNNs) are the state-of-the-art tools for automatic segmentation of 3D echocardiography images, and recent studies have explored their application to cardiac catheter localization. Yang et al. applied a 2D CNN architecture for catheter segmentation in 3D TEE by stacking together adjacent 2D axial slices [8]. To mitigate the possible loss of 3D semantic information after slicing, they subsequently introduced image patch extraction and multi-planar slicing [9]. Although the patch-based approach benefits from speed and low GPU memory usage, it may limit the network's ability to capture the full contextual information of the image. To address this, Yang et al. adopted a hybrid loss function that integrates voxel-wise loss with contextual loss for 3D UNet training, encouraging the network to learn a better contextual representation [10,11]. Alternatively, full 3D image information has been leveraged by combining 3D encoder and projection layers for features dimension reduction along axial and lateral dimensions, achieving efficient catheter segmentation in 3D TEE [12]. Despite these advancements, automatic catheter segmentation still requires additional complex post-processing to extract meaningful clinical information, such as catheter position and orientation. Furthermore, CNN methods for semantic segmentation rely on accurate ground truth (GT) annotations, which are difficult to obtain due to the noise and artifacts present in TEE images when visualizing metallic catheters. This frequently leads to over-segmentation of the catheter, producing numerous false positives that negatively affect the localization [13]. Consequently, these approaches have been applied to detect intracardiac catheters with simple shapes, such as ablation catheters or guide wires that are typically characterized by a straight configuration during the procedure.

1.2. Main contribution

In this work, we present the first automated approach for detecting, localizing, and characterizing the clip on the tip of the MitraClip delivery catheter. For clarity, we will refer to the MitraClip device, specifically the component involved in the grasping of MV leaflets during the procedure, simply as the clip. Unlike previous efforts that focused solely on catheter segmentation, our method is designed to accurately extract the clip's spatial features, which are essential for intraoperative guidance during TEER. We introduced a novel deep learning-based framework for clip detection in volumetric echocardiography data. Beyond simple segmentation, our approach enabled a comprehensive characterization of clip orientation and positioning, providing key parameters to support real-time surgical decision-making. Our method reduces the reliance on manual slice selection and expert interpretation, allowing for standardized and efficient intraoperative visualization. By addressing these challenges, our approach has the potential to enhance procedural accuracy, reduce operative time, and improve overall patient outcomes in TEER interventions.

2. Methods

2.1. Dataset collection

To implement and validate our proposed clip detection method, we collected 4D TEE recordings in a heart simulator designed to replicate transcatheter MV repair procedures *in vitro*. The simulator included anatomically realistic phantoms of:

- esophagus
- access veins, i.e., femoral, iliac, and hepatic veins as well as the inferior vena cava connected to the right atrium
- heart, including the relevant intracardiac structures, i.e., interatrial septum, left atrium, and MV with moving leaflets

All phantoms were immersed in demineralized water to enable TEE imaging. The setup was complemented by a MitraClip system (Abbot, G4 version) with a XTW clip and by a GE Vivid E95 scanner (GE Vingmed Ultrasound, Horten, Norway) with 4Vc-D TEE probe. At the beginning of experiments, the TEE probe was advanced next to the heart phantom through the esophagus phantom. It was then kept the probe at midesophageal level, ensuring procedural views of both MV and the MitraClip catheter in 2D and 3D modes. The MitraClip catheter was inserted in the phantom of the right femoral vein and driven through the interatrial septum and into the left heart (Fig. 1). We followed the procedural guidelines for TEER with the MitraClip system [4,5] for positioning the clip above the MV. The steps for positioning, which are preparatory for the deployment of the clip in either the central, lateral, or medial portion of the MV, included:

1. Advancing the clip delivery system through the sheath into the LA.
2. Steering the sheath catheter and simultaneously medially deflecting the clip delivery system while retracting the whole system to position the clip delivery system above the MV. This step included gentle anterior rotation of the clip delivery catheter accompanied by simple lateral rotations.
3. Adjusting the medial-lateral clip position and opening the clip.

The clip positioning procedure was repeated 44 times. For each procedure, TEE imaging was acquired at five stages: (i) before deflecting the clip delivery system, (ii) after positioning the system above the MV, and (iii–v) at three different clip opening angles. The final position of the clip relative to the MV (medial, central, or lateral) was varied across procedures. This resulted in a total of 220 4D TEE acquisitions. From each recording, a single frame corresponding to a 3D TEE image was extracted. Due to significant artifacts that prevented accurate

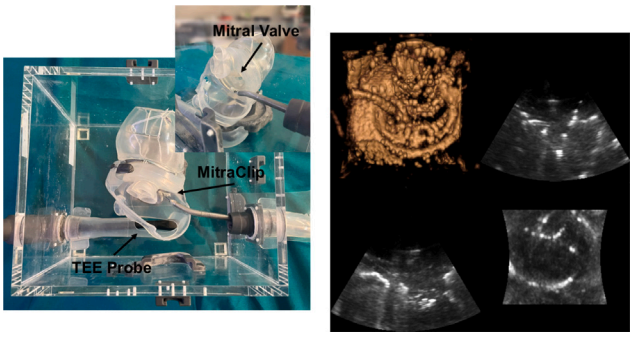

Fig. 1. Left: *in-vitro* setup for TEE acquisitions in TEER procedures. Black arrows indicate the TEE probe, the MitraClip and the MV. Right: example of *In-vitro* 3D TEE acquisition with MitraClip catheter.

annotation of the clip, 24 images were discarded. The remaining 196 3D TEE images were exported in Cartesian format and sampled with a uniform spacing of 0.5 mm, resulting in a mean size of $199 \times 248 \times 248$. This experimental design ensured that a sufficient number of images were collected to cover a wide range of clip positions, orientations, and opening angles.

2.2. Dataset annotation

2.2.1. Clip manual annotation

Two trained users segmented the MitraClip delivery catheter and clip, both assigned with the same label, using 3D Slicer [14]. Voxel thresholding was applied using the Otsu method [15], and the threshold intensity range was adjusted to isolate voxels corresponding to the MitraClip catheter and heart structures from the background. Using the 3D-rendered visualization of the segmentation, the users manually refined the segmentation by cutting away areas that were not considered part of the MitraClip catheter.

2.2.2. Design of MitraClip template models

A real XTW clip, detached from its delivery catheter, was scanned using a 3D scanner (Artec Micro II, Artec 3D Technology, Senningerberg, Luxembourg) to digitize the object. The clip was scanned in three configurations: fully closed and open with opening angle (θ) equal to 60° and 120°, respectively. Scanned models were exported in *.stl* format and were used as references to sketch the template models of the clip through computer-aided design (CAD) modeling in Autodesk Fusion (Autodesk Inc., San Francisco, California, United States). We started designing the MitraClip model in closed configuration (Fig. 2, top row, right). The model was conceived as a two-part 3D object, and we exploited the longitudinal symmetry of the clip by creating only one half of it. The complete model was then obtained by mirroring the generated half. The first part represented the bottom portion of the clip, extending from the tip to the origin of the clip's arms begin (Fig. 2, top row, in red). The upper part consisted of the clip's arm (Fig. 2, top row, in green). The bottom part was created using the loft method by fitting three elliptical half sections sketched along the clip axial direction at 1 mm, 3 mm, and 6 mm from the tip. The dimensions of these sections were chosen to match the corresponding segments of the scanned model. Similarly, the upper part was created by fitting two elliptical half sections sketched at 6 mm and 18 mm from the clip's tip and fitted using the loft method. Then, nine template models were automatically generated by rigidly tilting the arm. These were characterized by θ ranging from 10° up to 90° (10° step). To account for the actual deformation of the real object and ensure the accuracy of the templates, for the templates with θ equal to 60° and 120° the length and the cross-sectional dimensions of the arm were adjusted to match their values measured in the corresponding scanned models.

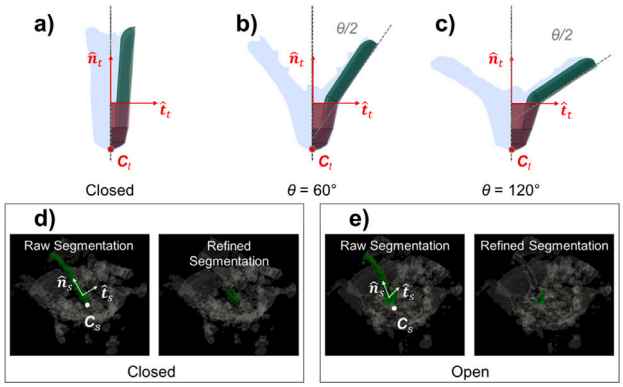

Fig. 2. (a)–(c): CAD models of the clip in three different configurations. Only one longitudinal half of the CAD model is shown with two different color for the bottom part (in red) and the upper part (in green). 3D representations of the scanned real clip are shown in transparency at the corresponding configurations. The clip tip ($\mathbf{C}_t$), longitudinal axis ($\hat{\mathbf{n}}_t$) and transversal axis ($\hat{\mathbf{t}}_t$) are highlighted in red for each template. (d)–(e): Two examples of the clip GT segmentations refinement using the template models in closed and open configurations. For each example, the raw segmentation mask is depicted on the left hand side, including its clip tip ($\mathbf{C}_s$), longitudinal axis ($\hat{\mathbf{n}}_s$) and transversal axis ($\hat{\mathbf{t}}_s$), while the refined mask is shown on the right hand side.

These were linearly interpolated to obtain the length and the cross-sectional dimensions of the arm in the other open configurations, thus representing the gradual changes in clip morphology across its range of configurations.

2.2.3. Clip annotation CAD-based refinement

To obtain accurate GT data, manual segmentations were automatically refined using the clip templates generated through CAD modeling. To this aim, given a manual segmentation mask, the following steps were implemented:

1. the manual segmentation mask was converted into a triangulated surface (Ω_s) by marching cubes algorithm;
2. the extremities of (Ω_s) were identified as the vertices associated with the minimum and maximum value of the first eigenvalue of the Laplace–Beltrami (LB) operator [16];
3. the extremity closest to the center of the image, and hence to the MV, was identified as the tip of the clip ($\mathbf{C}_s$);
4. the longitudinal and transversal axes ($\hat{\mathbf{n}}_s$ and $\hat{\mathbf{t}}_s$) of Ω_s were identified as the first and second principal direction, respectively, yielded by principal component analysis (PCA) on the 3D coordinates of the vertices of Ω_s;
5. each template (T_i), characterized by clip tip ($\mathbf{C}_{ti}$), longitudinal axis ($\hat{\mathbf{n}}_{ti}$), and transversal axis ($\hat{\mathbf{t}}_{ti}$), was rigidly co-registered on the segmentation mask so to have $\mathbf{C}_{ti}$ superimposed to $\mathbf{C}_s$, and $\hat{\mathbf{n}}_{ti}$ and $\hat{\mathbf{t}}_{ti}$ aligned with $\hat{\mathbf{n}}_s$ and $\hat{\mathbf{t}}_s$, respectively;
6. the Dice score between each co-registered template T_i and the segmentation mask was computed;
7. the co-registered template with maximal Dice score was used to refine the raw segmentation mask.

Figs. 2d and 2e, show two examples of GT segmentation before and after refinement through template registration for clips in closed and open configurations, respectively.

2.3. Application pipeline

We developed an automated three-stage pipeline for clip detection from 3D TEE images (Fig. 3). First, a 3D CNN based on a UNet architecture is employed to automatically segment the clip (Fig. 3, orange

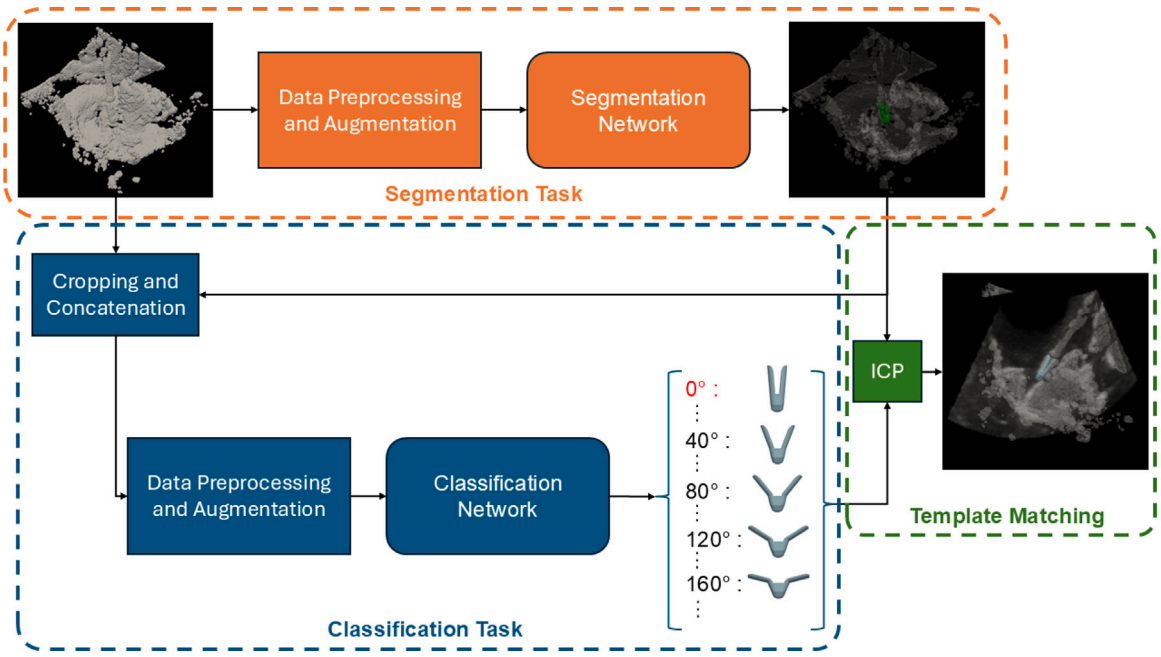

Fig. 3. Schematic representation of the proposed pipeline. Orange box: **Segmentation Task** — The clip is segmented from 3D TEE image. Blue box: Classification Task — The input 3D TEE image is cropped around the clip using the segmentation output as reference. The cropped image is then concatenated channel-wise with the segmentation mask and passed to the classification network to predict the clip configuration. Green Box: **Template Matching** — A clip template is selected based on the predicted configuration and rigidly registered to the segmentation output. The proposed pipeline enables automatic slicing of the 3D TEE volume and quantification of the clip status.

box). The largest connected component is extracted from the predicted segmentation to rule out possible spurious unconnected regions. Based on the resulting mask, images are cropped. Second, another 3D CNN processes the cropped image together with the predicted segmentation mask, concatenated channel-wise, to classify the clip configuration (Fig. 3, blue box). Third, a template corresponding to the predicted clip configuration is rigidly registered to the output of the 3D segmentation model (Fig. 3, green box).

2.3.1. Segmentation and classification steps

(a) Neural network architectures

For the segmentation task, four CNN architectures were considered:

- UNet [17] with five resolution levels, each of them consisting of a double convolutional block, resulting in approximately 20M trainable parameters.
- SegResNet [18] with asymmetrical design. The encoder was larger, with four down-sampling stages characterized by 1, 2, 2, and 4 ResNet blocks [19], respectively. These were aimed to enhanced feature extraction. The decoder to reconstruct the segmentation mask was more compact. This architecture resulted in approximately 18 millions of trainable parameters.
- Attention UNet [20] with five resolution levels, each one consisting of a double convolutional block, and attention gates to refine feature maps by focusing on salient regions, enabling the network to prioritize relevant areas for segmentation [21]. This architecture was characterized by approximately 24M trainable parameters.
- UNetR [22] integrating a transformer-based encoder into the UNet framework, featuring five resolution levels and twelve attention heads. This hybrid design allowed the network to capture global, multi-scale information by learning sequential representations of the input volume [22]. As expected, the UNetR was the heaviest architecture, with the number of trainable parameters exceeding 100M.

For the classification task, we utilized ResNet-50 [19] and DenseNet [23], two widely used CNN architectures designed for image classification. Both architectures leverage skip connections to improve gradient flow during training, enabling faster convergence and better performance on complex classification tasks. ResNet-50 achieves this through its residual blocks, while DenseNet connects each layer to every subsequent layer within the same block, facilitating feature reuse and efficiency in parameter utilization.

(b) Implementation details

For both the segmentation and classification networks, the dataset was split into training, validation, and testing subsets using a 70:10:20 ratio. Images were resampled to achieve a uniform voxel spacing of 1 mm in all directions. Clip configuration classes were imbalanced within the dataset. To ensure balanced representation, we employed a weighted random sampler during data loading for both segmentation and classification network training, based on configuration frequencies. This approach ensured an even distribution across classes within each batch for both segmentation and classification tasks. In training only, data augmentation was performed on-the-fly, including intensity transforms, i.e., intensity scaling and random Gaussian noise, as well as spatial transforms, i.e., random rotation, random axis flip, and random elastic deformation.

During the training of the segmentation network, patches of $128 \times 128 \times 128$ voxels were randomly extracted from the input 3D images. The segmentation networks were trained over 500 epochs with a batch size of 8, using the Adam optimizer, with the learning rate initially set to 0.001 and dynamically adjusted using a cosine annealing scheduler. A weighted combination of Dice and Focal losses [24], with weights of 0.6 and 0.4, respectively, was minimized during the training phase. A 0.1 dropout rate, determined through hyper-parameter search, was selected as it provided the best results in early experiments.

Once the segmentation network training had converged, the best-performing segmentation model on the validation set, comprising 19 images, was used to extract homogeneous patches of $30 \times 30 \times 30$ voxels from the input images, centered on the segmentation regions.

Table 1
Distribution of clip configurations in the training and test sets.

Angle Configuration	0°	20°	40°	60°	80°	100°	120°	140°	160°	180°
Training Set (156)	69	2	6	11	11	15	18	7	2	15
Test Set (40)	19	0	1	2	3	5	3	1	1	5

Table 2
Segmentation performance across the test set. Average Dice score, average surface distance (ASD) and 95% Hausdorff distance (HD95) are reported for the UNet, AttentionUNet, SegResNet and UNetR. The best value for each segmentation metric across all models is highlighted in bold.

Metric	UNet	Attention UNet	SegResNet	UNetR
Dice score	0.59 ± 0.16	**0.62 ± 0.14**	0.53 ± 0.12	0.53 ± 0.19
ASD (mm)	0.81 ± 1.00	**0.76 ± 1.03**	1.04 ± 1.10	1.32 ± 3.93
95% HD (mm)	2.73 ± 2.88	**2.44 ± 2.40**	3.88 ± 2.91	4.35 ± 5.18

Each patch was concatenated channel-wise with the corresponding segmentation mask of the same dimensions and used as input to train the classification networks. The classification models were trained over 600 epochs with a batch size of 40, using the Adam optimizer. The initial learning rate was set to 0.00001 and halved every 300 epochs. Cross-entropy loss was used to train the classification networks.

The best models on the validation set were saved for inference on the test set for both the segmentation and the classification networks. All models were implemented using PyTorch [25] and the MONAI library [26]. Training and inference were performed on an NVIDIA RTX 4090 GPU with 24 GB of memory.

2.3.2. Template matching step

Segmentation masks yielded by the first step of the pipeline were refined through a two-step approach. First, the CNN-based segmentation mask of the clip was converted into a 3D surface using the marching cubes algorithm. Second, the surface of the clip template, selected based on the classification performed in the second step of the pipeline, was aligned to the surface of the predicted segmentation mask through the iterative closest point (ICP) algorithm [27]. Specifically, we used the ICP implementation from the VTK library [28]. To balance accuracy and computational efficiency, the maximum number of iterations was set to 1000.

2.4. Evaluation

The test set consisted of 40 3D TEE images displaying the clip in several configurations. Clip configurations and number of 3D TEE images for configuration within the test set are shown in Table 1. To demonstrate the efficacy of the proposed pipeline for clip detection, we separately evaluated the segmentation and classification performances. For the segmentation task, the predicted segmentations were compared with the refined clip GTs described in Section 2.2.3. The largest connected component was extracted from the predicted segmentations before computing the evaluation metrics. Standard segmentation metrics were computed, including the Dice score, average surface distance (ASD), and 95% Hausdorff Distance (95% HD). For the classification task, we evaluated the performance by computing precision, recall, and F1-score for each class in the test set, as well as the weighted average performance across all classes defined by

$$\text{Weighted Average} = \sum_{i=1}^{N} \frac{n_i}{N_{\text{total}}} \cdot M_i$$

where n_i is the number of true instances for class i, M_i is the metric value (e.g., precision, recall, or F1-score) for class i, and $N_{\text{total}} = \sum_{i=1}^{N} n_i$ is the total number of instances across all classes. Using the best classification architecture between ResNet-50 and DenseNet, we recomputed the segmentation performance after registering the predicted template based on the classification network's output. Upon verifying the normal distribution of data through a Shapiro–Wilk normality test, a two-sample t-test was run to assess the improvement in segmentation following registration with the template. Differences were deemed statistically significant for $p < 0.05$.

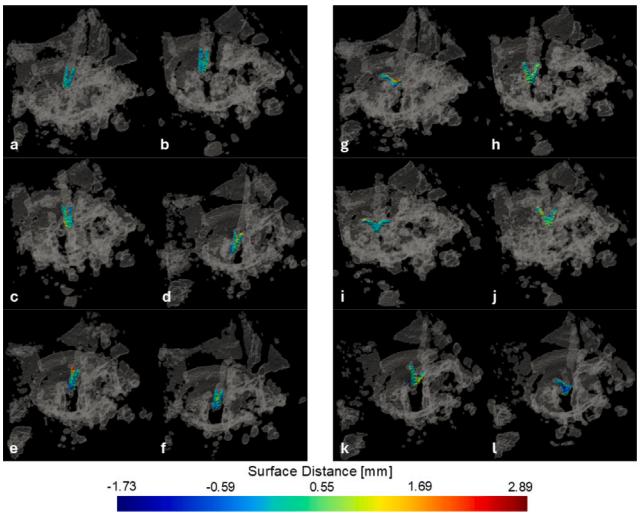

Fig. 4. Segmentation examples provided by the Attention UNet for closed (a–f) and open clip configurations (g–l). Surface distance heatmaps computed respect to the GT, are overlaid on the segmentation masks.

3. Results

3.1. Segmentation performance

Table 2 summarizes the average performance across the test set achieved by the network architectures considered for clip segmentation. Among all the models, Attention UNet demonstrated the best performance achieving a Dice score of 0.62 ± 0.14, ASD of 0.76 ± 1.03 mm and 95% HD 2.44 ± 2.40 mm. The simple UNet achieved comparable performance but exhibited a lower Dice score and higher distance metrics compared to Attention UNet. In contrast, SegResNet and UNetR showed weaker results, with Dice score below 0.6, ASD around 1 mm, and 95% HD values close to 4 mm.

Fig. 4 provides examples of clip segmentations obtained using the Attention UNet. It depicts the device in a closed configuration (Fig. 4, a–f) and in several open configurations (Fig. 4, g–l). Overall, the segmentation network successfully localized the clip and captured its shape. The bottom part of the clip was the least challenging to detect, as indicated by lower distance errors vs. GTs. Conversely, the upper part proved more challenging, yielding higher distance errors, especially in open configurations where one of the clip's arms was occasionally missed (Fig. 4, g and l).

3.2. Classification performance

Table 3 presents the classification performance achieved by DenseNet and ResNet-50 across the test set for cropped and uncropped input data according to the segmentation output. For this evaluation, the Attention UNet was used as the segmentation network, as it achieved the best performance in the segmentation task.

Both DenseNet and ResNet-50 performed significantly better when cropped input data were used in combination with channel-wise concatenation of the segmentation mask, reaching weighted average F1-scores of 0.80 and 0.75, respectively, across the different clip configurations. DenseNet slightly outperformed ResNet-50 in precision (0.80 vs. 0.77) and achieved higher recall (0.83 vs. 0.75). For the 0° and 180° configurations, DenseNet reached optimal precision, recall, and F1-scores (F1 = 1.0 and 0.9, respectively). Worse results were obtained for 80° configurations, achieving F1-scores of 0.34. The model performed

Table 3
Classification performance of DenseNet and ResNet-50 for clip configurations across the test set. Results are reported for cropped and uncropped input data. Metrics include precision, recall, and F1-score for each configuration, along with the corresponding weighted averages. The cropped input uses segmentation-assisted preprocessing, while the uncropped input relies on randomly extracted patches. *Classification performance for 20° configuration is consistently zero because this class is absent from the test set.

	Cropping & Concatenation						No cropping						
	DenseNet			ResNet-50			DenseNet			ResNet-50			
Configuration	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Support
0°	1.0	1.0	1.0	1.0	0.78	0.88	0.25	0.05	0.09	0.0	0.0	0.0	19
20°*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
40°	1.0	1.0	1.0	0.0	0.0	0.0	0.08	1.0	0.15	0.0	0.0	0.0	1
60°	1.0	0.5	0.67	0.67	1.0	0.88	0.2	0.5	0.28	0.0	0.0	0.0	2
80°	0.34	0.34	0.34	0.0	0.0	0.0	0.0	0.0	0.0	0.03	0.03	0.01	3
100°	0.6	0.6	0.6	0.57	0.8	0.67	0.0	0.0	0.0	0.0	0.0	0.0	5
120°	0.6	0.1	0.75	0.75	1.0	0.86	0.0	0.0	0.0	0.0	0.0	0.0	3
140°	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1
160°	0.0	0.0	0.0	0.34	1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	1
180°	0.84	1.0	0.9	1.0	1.0	1.0	0.14	0.2	0.17	0.0	0.0	0.0	5
Weighted Average	0.80	0.83	0.80	0.77	0.75	0.75	0.15	0.1	0.08	0.51	0.15	0.17	40

poorly on low-support classes, such as 140° and 160°, missing the sole sample of each in the test set. ResNet-50 excelled in classifying 0° (F1 = 0.88), 60° (F1 = 0.88) and 120° (F1 = 0.86) configurations, but underperformed in configurations with limited supports, such as 40°, 80°, and 140°.

For uncropped input data, both DenseNet and ResNet-50 struggled significantly, achieving a weighted average F1 score of 0.08 and 0.17, respectively. The models exhibited near-zero precision and recall across most configurations, failing to accurately classify configurations such as 0°, 60°, and 100°.

Fig. 5 shows the confusion matrices for the classification performances achieved by DenseNet and ResNet-50 across the test set for cropped and uncropped input data. The confusion matrices confirm the previous observations, showing significantly better performance with cropped input data for both DenseNet and ResNet-50. The matrices for cropped input are more diagonally dominant, reflecting a higher rate of correct predictions. DenseNet's confusion matrix for cropped input shows less off-diagonal entries than the confusion matrix of ResNet-50, indicating more consistent classification. Misclassifications primarily resulted in errors of one configuration step (20°), particularly for classes such as 80°, 100°, and 160°. In contrast, with uncropped input, the confusion matrices for both models display substantial off-diagonal entries for both DenseNet and ResNet-50, reflecting significant misclassifications. Predictions are often biased toward dominant configurations, such as 0° and 180°, and errors are distributed across the matrix.

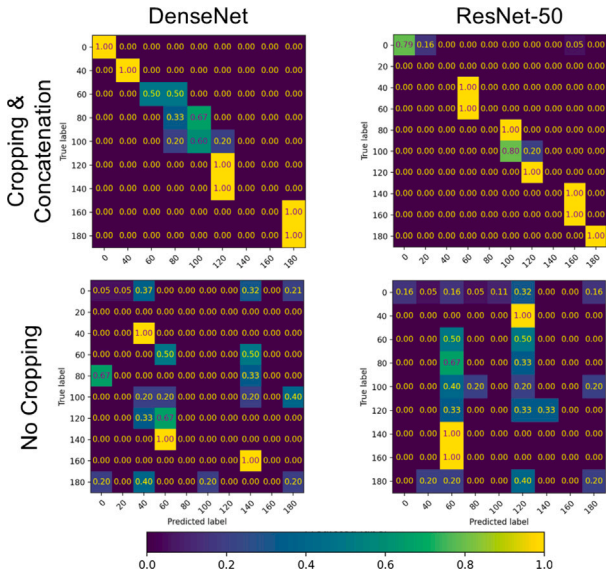

Fig. 5. Confusion matrices for DenseNet and ResNet-50 classification performance across the test set. Matrices are shown for both cropped and uncropped input data. Normalized predictions are shown in each matrices.

3.3. Segmentation performance after refinement

Table 4 summarizes the average performance across the test set achieved by the network architectures after template registration through the ICP algorithm. Templates were selected based on the classification outputs provided by the DenseNet, which achieved the best performance in the classification task. For this step, DenseNet received input data cropped based on the segmentation output of the evaluated network and concatenated channel-wise with the corresponding segmentation mask.

Similar to the raw segmentation performance reported previously, Attention UNet demonstrated the best performance achieving a Dice score of 0.59 ± 0.17, ASD of 0.69 ± 0.99 mm and 95% HD 1.83 ± 1.63 mm. UNet achieved larger distance metrics, with an ASD of 0.76 ± 1.14 mm and a 95% HD of 2.10 ± 2.54 mm, and showed a slightly lower average Dice score of 0.57 ± 0.17 compared to Attention UNet. In contrast, SegResNet and UNetR slightly underperformed as compared to the other architectures, achieving a Dice score slightly above 0.5, on average. UNetR showed the weakest overall performance, with an average ASD value and the average 95% HD value above 1 mm and 2.50 mm, respectively.

Compared to the raw segmentation results (Table 2), template registration using the ICP algorithm notably improved ASD and 95% HD values. For example, SegResNet showed the most notable improvement in ASD, which decreased from 1.04 ± 1.10 mm to 0.89 ± 1.08 mm, and in 95% HD, which decreased from 3.88 ± 2.91 mm to 1.35 ± 3.92 mm. Similarly, UNetR exhibited a significant reduction in 95% HD, improving from 4.35 ± 5.18 mm to 2.99 ± 5.02 mm. However, the Dice score for all segmentation networks did not improve; rather, it slightly worsened after refinement through template registration.

Fig. 6 compares the segmentation performance of Attention UNet across three scenarios: the raw segmentation output, the refined segmentation output based on the predicted classification provided by DenseNet, and the refined segmentation output based on the actual classification. The Dice score slightly worsened after segmentation refinement, with the median value decreasing from 0.62 to 0.55. In contrast, ASD and 95% HD improved, with median values decreasing from 0.63 mm to 0.59 mm and from 2.0 mm to 1.45 mm, respectively. Segmentation performance based on the actual classification yielded median values of 0.57, 0.56 mm, and 1.41 mm, respectively, which were comparable to those obtained using the predicted classification.

Fig. 7 provides qualitative examples of clip segmentations obtained with Attention UNet after refinement through template matching. It

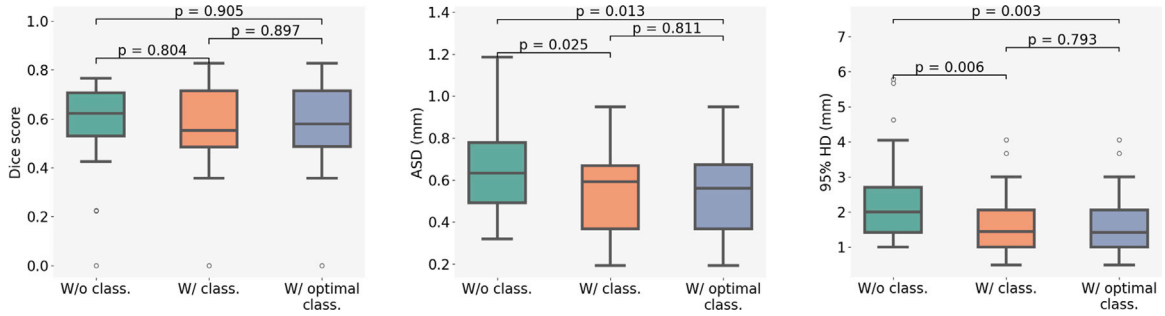

Fig. 6. Boxplots for segmentation performance achieved by the Attention UNet with and without segmentation refinement through template matching. For Dice score, ASD and 95% HD the boxplots compare performance metric for the raw segmentation output (W/o classification, green box), the refined segmentation output based on the predicted configuration from DenseNet (W/ classification, orange box) and the refined segmentation based on the actual configuration (W/ optimal classification, purple box). Level of significance is reported above the box plot respect to the performance metric for the raw segmentation output.

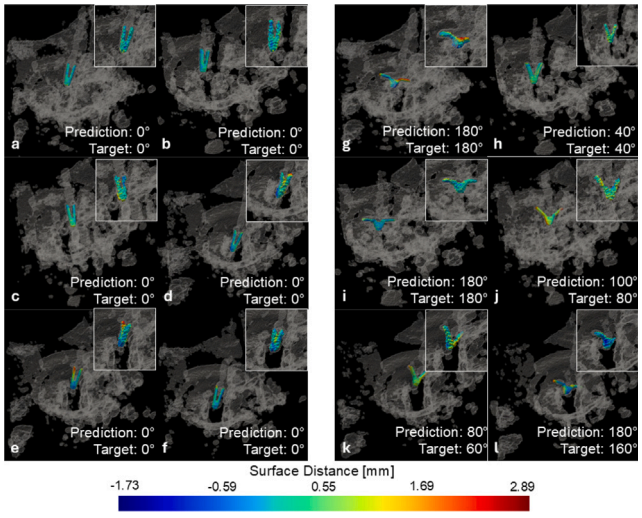

Fig. 7. Segmentation examples provided by Attention UNet for closed configurations (a–f) and open clip configurations (g–l). Each case includes a zoomed view of the raw segmentation and reports the predicted configuration from DenseNet alongside the target configuration used for template selection. Surface distance heatmaps, computed relative to the GT, are overlaid on the segmentation masks.

depicts the device in a closed configuration (Fig. 7, a–f) and in several open configurations (Fig. 7, g–l). Each case shows a zoomed view of the raw segmentation and reports the predicted configuration from DenseNet alongside the target configuration used for template selection. Overall, template matching improved segmentation outputs by enhancing consistency with GTs, as reflected in improved distance metrics. Template matching successfully restored small details, such as the clip arms, that were sometimes lost in raw segmentation outputs, particularly in open configurations (Fig. 7, g and l). Incorrect classification predictions, and consequently incorrect template matching, led to slight worsening of distance metrics (Fig. 7, j, k and l). However, these errors were negligible, as they involved small differences in clip configurations. These variations are indiscernible given the resolution and image quality of echocardiography and are not significant in the context of intraprocedural guidance.

3.4. Influence of clip configuration

Tables 5 and 6 report the average segmentation performance of the Attention UNet across the test set, stratified by the relative clip position with respect to the MV orifice, both before and after segmentation refinement. Both tables highlight better performance when the clip is positioned centrally or laterally, with Dice scores greater than 0.6, ASD below 0.50 mm, and 95% HD below 1.45 mm after segmentation refinement.

Table 4
Segmentation performance across the test set after segmentation refinement through template matching. Template selection was based on predictions from DenseNet. Average Dice score, average surface distance (ASD) and 95% Hausdorff distance (HD95) are reported for the UNet, AttentionUNet, SegResNet and UNetR. The best value for each segmentation metric across all models is highlighted in bold.

Metric	UNet	Attention UNet	SegResNet	UNetR
Dice score	0.58 ± 0.16	**0.59 ± 0.17**	0.50 ± 0.18	0.52 ± 0.20
ASD (mm)	0.76 ± 1.14	**0.69 ± 0.99**	0.89 ± 1.08	1.35 ± 3.92
95% HD (mm)	2.10 ± 2.54	**1.83 ± 1.63**	2.37 ± 2.02	2.99 ± 5.02

Table 5
Segmentation performance across the test set for medial, central, and lateral clip positions relative to the MV, before segmentation refinement. Average Dice score, ASD, and 95% HD are reported for the Attention UNet. The best value for each metric is highlighted in bold.

Attention UNet Position	Medial	Central	Lateral
Dice score	0.57 ± 0.18	**0.65 ± 0.11**	0.63 ± 0.12
ASD (mm)	1.13 ± 1.78	**0.57 ± 0.18**	0.63 ± 0.26
95% HD (mm)	3.85 ± 3.73	1.83 ± 0.99	**1.77 ± 0.79**

Table 6
Segmentation performance across the test set for medial, central, and lateral clip positions relative to the MV after segmentation refinement through template matching. Average Dice score, average surface distance (ASD) and 95% Hausdorff distance (HD95) are reported for the Attention UNet. The best value for each metric is highlighted in bold.

Attention UNet Position	Medial	Central	Lateral
Dice score	0.48 ± 0.20	**0.64 ± 0.11**	**0.64 ± 0.13**
ASD (mm)	1.38 ± 1.90	0.50 ± 0.23	**0.47 ± 0.22**
95% HD (mm)	3.68 ± 4.02	1.43 ± 0.70	**1.34 ± 0.69**

3.5. Inference time

The proposed pipeline takes 2.55±0.58 s to process a 3D TEE volume and detect the clip within it. The required time is subdivided as follows: 0.01±0.01 s for the segmentation task, 0.09±0.03 s for the classification task and 2.44±0.58 s for model refinement through template matching. The notable time-expense of the final step results from leveraging an iterative algorithm, i.e., ICP: its time-expense can increase massively depending on the number of points considered for surface alignment, the convergence threshold, and the maximum number of iterations allowed for convergence.

4. Discussion

We presented an automated pipeline for clip detection from intraoperative 3D TEE images, enabling accurate localization of the device, configuration quantification, and enhanced 3D visualization within the intracardiac space. To the best of our knowledge, this is the first attempt to automate the detection of a device of such complexity from intraoperative 3D TEE images. The pipeline offers two significant advancements for MitraClip procedures. First, it provides a fast detection of the clip, enabling localization in the intracardiac space for the operator. Second, it provides a fast classification of the clip pose, hence essential information for device orientation and position. Of note, this information could be combined with the result of automated MV segmentation and reconstruction to yield quantitative information on the pose of the clip relative to the target region.

We trained and evaluated four different 3D CNN architectures for clip segmentation: UNet, Attention UNet, SegResNet, and UNetR. While all these models share a common backbone structure based on the traditional UNet's encoding–decoding architecture, each of them incorporates distinctive features tailored to address specific challenges in medical imaging, including echocardiography. These architectures are well-established for medical image segmentation. Traditional UNet [29] and its custom variants [8–12] have been widely applied to catheter segmentation in echocardiography, while Attention UNet [21] and SegResNet [30] have been successfully used for MV segmentation from 2D and 3D TEE. UNetR incorporates multi-head attention mechanisms typical of Transformer architectures [31], which have shown promising performance in segmentation tasks beyond their original use in natural language processing. These networks were trained using manually generated GT segmentations, refined by matching with CAD-based template models of a real MitraClip device. This refinement step ensured that the GT segmentations preserved the actual dimensions of the device, enhancing reliability and precision in detection. This approach was required because accurate manual delineation of implantable devices is particularly challenging in TEE images due to noise, artifacts, and the presence of metallic catheter components, which often lead to over-segmentation and false positives that can compromise localization accuracy [13]. Among the evaluated architectures, Attention UNet demonstrated the best segmentation performance, on average. As shown in Fig. 4, Attention UNet reliably segmented the clip, accurately reconstructing the clip's tip and arms with minimal error vs. GTs in closed configuration (Fig. 4 a–f). However, in open configurations, the network occasionally failed to fully reconstruct the clip, missing one of the two arms. This typically occurred when the clip was positioned near heart structures, where the limited image contrast hindered clear clip visualization (Fig. 4 g and l). Attention UNet benefited from its attention gates, which enabled better localization of salient features [20], improving segmentation accuracy for challenging tasks like clip detection, where the device occupies a small region within the broader field of view in 3D TEE. This architecture guaranteed state-of-the-art performance for cardiac catheter detection from 3D TEE, comparable to previous studies [9–12,29] that employed UNet-based architectures and reported an average Dice score ranging between 0.57 and 0.7, and average HD between 7 and 1 voxels. Unlike those studies, our approach did not rely on additional preprocessing nor on architecture enhancements, such as multi-planar slicing [8], patch-of-interest extraction [10], contextual information enhancement [9,12], or spatial complexity reduction using projection layers [11] in UNet's decoder path. Furthermore, previous studies focused on segmenting simpler structures, such as ablation catheters [29] or guide wires [9–12], which typically maintain a straight configuration during procedures. In contrast, the MitraClip system includes a highly steerable delivery catheter and a clip with dynamically adjustable arms. Hence, it presents unique challenges requiring more sophisticated detection and segmentation strategies.

For the classification task, we compared the performance of DenseNet and ResNet-50. Both architectures benefited from segmentation-assisted input cropping and channel-wise concatenation with the segmentation mask, which allowed the classifier to focus on features extracted from the region of interest provided by the segmentation network. This approach improved the classification accuracy without requiring additional operations, as the segmentation network was already part of the pipeline. Using cropped input data, DenseNet outperformed ResNet-50, achieving a higher weighted average F1-score (0.80 vs. 0.75). The dense connectivity of DenseNet likely contributed to more robust feature reuse, resulting in superior classification accuracy. Optimal precision, recall, and F1-scores were achieved for some configurations, particularly closed and 180°. However, DenseNet struggled to classify configurations with limited support, such as 140° and 160°. Despite these challenges, DenseNet's errors were generally minor, with most misclassifications differing by only 20° (Fig. 5). These small misclassification errors are unlikely to impact clinical outcomes significantly, as they do not hinder the overall understanding of the clip configuration.

In the final step of our proposed pipeline, template registration was employed based on the predicted configuration to enhance segmentation accuracy. For simpler catheter shapes, such as ablation catheters or guide wires that maintain a mostly straight configuration, model fitting was previously proposed to improve segmentation output [8]. However, the higher geometrical complexity of the clip and its broad range of configurations required template matching instead. This approach proved effective in improving distance metrics. For instance, SegResNet's ASD decreased from 1.04 ± 1.10 to 0.89 ± 1.08 mm and the 95% HD decreased from 3.88 ± 2.91 to 1.35 ± 3.92 mm. Similarly, UNetR exhibited a reduction in 95% HD from 4.35 ± 5.18 mm to 2.99 ± 5.02 mm. Unexpectedly, for all architectures, the average Dice score slightly worsened after refinement. However, as shown for Attention UNet, this decrease was not statistically significant ($p > 0.05$, Fig. 6). Conversely, improvements in distance metrics, which better reflect structural alignment, were more pronounced. Although not all differences were statistically significant, this trend suggests that the limited dataset size may have influenced the analysis. Notably, Attention UNet achieved median ASD values of 0.59 mm and 0.56 mm, and median 95% HD values of 1.45 mm and 1.41 mm, respectively, when using the predicted and actual classifications. These findings strongly suggest that the observed misclassifications of clip configuration negligibly contributed to distance errors and did not significantly affect the accuracy of clip reconstruction vs. GTs. Overall, template matching was effective in restoring structural details and the original shape of the detected device, preserving key landmarks, such as clip tip and arms, to be detected in 3D TEE for intraprocedural guidance. This behavior explains the observed improvement in distance metrics, which are more sensitive to shape variations than overlap-based metrics like the Dice score. For closed configurations (Fig. 7, a–f), template matching accurately reconstructed the clip's arms, often missed in raw segmentations due to noise or low contrast. In open configurations, template matching improved alignment by restoring the original shape in cases where one arm was missed, enhancing agreement with GTs (Fig. 7, g and l). However, as template matching relies on the ICP algorithm, its effectiveness depends on the reliability of the target structure (in this case, the predicted segmentation) and the topological similarity between the template and the target structure. Incomplete segmentations or erroneous classifications can lead to suboptimal registration, potentially explaining the lack of Dice score improvement after segmentation refinement. Unlike distance metrics, the Dice score heavily relies on the degree of overlap between the predicted and target structures, making it less sensitive to cases where template matching restores shape details but fails to achieve optimal alignment with the GTs. Despite these subtleties, the pipeline would still enable automatic extraction of optimal 2D views to visualize the clip, since these could be positioned and oriented in the 3D TEE volume based on the pose of the reconstructed clip. Errors associated with template selection

were minor and did not compromise the accurate interpretation of the clip configuration. Ultimately, segmentation performance was found to vary according to the relative position of the clip with respect to the MV orifice, with higher accuracy observed when the device was positioned centrally or laterally. This may be explained by the fact that the delivery catheter, which brings the clip, is inserted through the atrial septum, a structure positioned medially to the MV. As a consequence, the medial portion of the MV is the most challenging to visualize in 3D TEE, as a portion of the catheter is almost always interposed between the probe and the target structures. This interposition generates imaging artifacts that hinder a clear distinction between the clip and the surrounding cardiac anatomy. It should be noted, however, that although the end position of the clip (medial, central, or lateral) varied across procedures during the dataset acquisition, this variation was not performed in a systematic or balanced manner, which may have influenced the observed differences in performance.

The proposed pipeline has the potential to provide accurate and reliable guidance during MitraClip procedures in near-real-time (2.55 ± 0.58 s). It enhances image acquisition and interpretation in intracardiac scenarios, which are typically hindered by echocardiographic limitations such as noise and low contrast. Moreover, our automated pipeline could be extended to other devices for TEER, e.g., the Pascal Precision System [32] by Edwards Lifesciences, or for transcatheter tricuspid valve repair, e.g., Triclip [33] by Abbott, with minimal modifications. This flexibility broadens its applicability and offers significant potential to improve procedural outcomes across various interventional cardiology applications.

4.1. Limitations and future works

This study utilized a dataset acquired *in vitro* on a single heart simulator with consistent anatomical phantoms and a single MitraClip size. While this controlled environment ensured consistent acquisitions and reliable ground-truth annotations, it inevitably limits the variability of the dataset and may reduce the generalizability of the trained neural network. In particular, the setup cannot capture differences in the relative position between the probe and the clip or the variation in patient tissue properties that affect ultrasound imaging. To comprehensively assess the performance of the method in real-world scenarios and mitigate potential overfitting, future research should focus on validation using more diverse datasets, including in-vivo data acquired from human subjects. Such data will enable the evaluation of the pipeline's robustness and generalization capability under clinical conditions, including variations in patient anatomy and imaging quality, and the assessment of its impact on the current clinical workflow. If successfully translated, the proposed pipeline has the potential to enhance procedural effectiveness, reduce procedural times, and mitigate the risk of adverse outcomes by providing real-time quantitative feedback to operators.

Another limitation lies in the dataset's imbalanced representation of clip configurations, which posed challenges for both model training and evaluation. While we initially aimed for a uniform sampling of opening angles, this was not technically feasible due to the mechanical constraints of the MitraClip delivery system. The device is actuated through a tendon-based mechanism controlled by manual knobs, which introduces non-linear and unpredictable behavior due to pre-tensioning and internal friction. As a result, consistently reproducing specific angles during data acquisition was not achievable. To mitigate the resulting imbalance, we applied oversampling during neural network training. While this reduced the impact of underrepresented configurations, classification errors, especially for intermediate angles, persisted. Although these errors were not critical for the final application, they indicate room for improvement. Future work could focus on acquiring a larger and more diverse dataset with a more balanced and controllable distribution of clip configurations, possibly through the development of a more precise actuation interface or synthetic data augmentation strategies. Notably, the model exhibited better performance in classifying fully closed and fully open configurations, likely due to their higher representation and more distinct visual features in 3D TEE volumes.

The segmentation and classification steps of the pipeline achieved fast inference times suitable for intraprocedural guidance. However, the segmentation refinement step, which relies on the ICP algorithm, significantly increased computational time. The ICP algorithm's computational intensity and dependence on the quality of segmentation output present challenges; imprecise segmentations can lead to suboptimal refinement. To address these challenges, a potential future direction could involve replacing the segmentation network with a model that directly infers the orientation and configuration of the device. End-to-end approaches for predicting 3D object orientation are well-established in general computer vision [34] and could be adapted for this application. Such an approach could eliminate the need for segmentation refinement, thereby streamlining the process.

5. Conclusion

We presented an automated pipeline for accurate clip detection and pose classification from intraprocedural 3D TEE images. The proposed method demonstrated high precision in device localization and reliable configuration assessment, addressing critical challenges in the real-time interpretation of echocardiographic data during MitraClip procedures. By simplifying and automating these tasks, this method has the potential to enhance procedural accuracy and efficiency, reducing the cognitive and physical workload on operators. Improved localization and configuration assessment directly support critical clinical decisions, potentially minimizing procedural times and mitigating the risks associated with adverse outcomes.

CRediT authorship contribution statement

Riccardo Munafò: Writing – review & editing, Writing – original draft, Software, Methodology, Conceptualization. **Simone Saitta:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Luca Vicentini:** Validation, Conceptualization. **Davide Tondi:** Software, Methodology. **Veronica Ruozzi:** Validation, Methodology. **Francesco Sturla:** Software, Methodology. **Giacomo Ingallina:** Validation, Conceptualization. **Andrea Guidotti:** Validation, Supervision. **Eustachio Agricola:** Supervision. **Emiliano Votta:** Writing – review & editing, Writing – original draft, Supervision.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used overleaf GPT in order to improve the readability and language of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

References

[1] Alec Vahanian, Friedhelm Beyersdorf, Fabien Praz, Milan Milojevic, Stephan Baldus, Johann Bauersachs, Davide Capodanno, Lenard Conradi, Michele De Bonis, Ruggero De Paulis, et al., 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), Eur. Heart J. 43 (7) (2022) 561–632.

[2] Ted Feldman, Saibal Kar, Sammy Elmariah, Steven C. Smart, Alfredo Trento, Robert J. Siegel, Patricia Apruzzese, Peter Fail, Michael J. Rinaldi, Richard W. Smalling, et al., Randomized comparison of percutaneous repair and surgery for mitral regurgitation: 5-year results of EVEREST II, J. Am. Coll. Cardiol. 66 (25) (2015) 2844–2854.

[3] Nicola Buzzatti, Mathias Van Hemelrijck, Paolo Denti, Stefania Ruggeri, Davide Schiavi, Iside Stella Scarfò, Diana Reser, Maurizio Taramasso, Alberto Weber, Giovanni La Canna, et al., Transcatheter or surgical repair for degenerative mitral regurgitation in elderly patients: a propensity-weighted analysis, J. Thorac. Cardiovasc. Surg. 158 (1) (2019) 86–94.

[4] M.A. Sherif, L. Paranskaya, S. Yuecel, S. Kische, O. Thiele, G. D'Ancona, A. Neuhausen-Abramkina, J. Ortak, H. Ince, A. Öner, MitraClip step by step; how to simplify the procedure, Neth. Hear. J. 25 (2017) 125–130.

[5] Charles B. Nyman, G. Burkhard Mackensen, Srdjan Jelacic, Stephen H. Little, Thomas W. Smith, Feroze Mahmood, Transcatheter mitral valve repair using the edge-to-edge clip, J. Am. Soc. Echocardiogr. 31 (4) (2018) 434–453.

[6] Riccardo Munafò, Simone Saitta, Giacomo Ingallina, Paolo Denti, Francesco Maisano, Eustachio Agricola, Alberto Redaelli, Emiliano Votta, A deep learning-based fully automated pipeline for regurgitant mitral valve anatomy analysis from 3D echocardiography, IEEE Access (2024).

[7] Hongxu Yang, Caifeng Shan, Alexander F. Kolen, Peter H.N. de With, Medical instrument detection in ultrasound: a review, Artif. Intell. Rev. 56 (5) (2023) 4363–4402.

[8] Hongxu Yang, Caifeng Shan, Alexander F. Kolen, Peter H.N. de With, Efficient catheter segmentation in 3D cardiac ultrasound using slice-based FCN with deep supervision and f-score loss, in: 2019 IEEE International Conference on Image Processing, ICIP, IEEE, 2019, pp. 260–264.

[9] Hongxu Yang, Caifeng Shan, Alexander F. Kolen, Peter H.N. de With, Improving catheter segmentation & localization in 3d cardiac ultrasound using direction-fused fcn, in: 2019 IEEE 16th International Symposium on Biomedical Imaging, ISBI 2019, IEEE, 2019, pp. 1122–1126.

[10] Hongxu Yang, Caifeng Shan, Tao Tan, Alexander F. Kolen, Peter H.N. de With, Transferring from ex-vivo to in-vivo: Instrument localization in 3d cardiac ultrasound using pyramid-unet with hybrid loss, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22, Springer, 2019, pp. 263–271.

[11] Hongxu Yang, Caifeng Shan, Arthur Bouwman, Alexander F. Kolen, Peter H.N. de With, Efficient and robust instrument segmentation in 3D ultrasound using patch-of-interest-FuseNet with hybrid loss, Med. Image Anal. 67 (2021) 101842.

[12] Hongxu Yang, Caifeng Shan, Alexander F. Kolen, Peter H.N. de With, Efficient medical instrument detection in 3D volumetric ultrasound data, IEEE Trans. Biomed. Eng. 68 (3) (2020) 1034–1043.

[13] Andre Mastmeyer, Guillaume Pernelle, Ruibin Ma, Lauren Barber, Tina Kapur, Accurate model-based segmentation of gynecologic brachytherapy catheter collections in MRI-images, Med. Image Anal. 42 (2017) 173–188.

[14] Ron Kikinis, Steve D. Pieper, Kirby G. Vosburgh, 3D slicer: a platform for subject-specific image analysis, visualization, and clinical support, in: Intraoperative Imaging and Image-Guided Therapy, Springer, 2013, pp. 277–289.

[15] Nobuyuki Ostu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1979) 62.

[16] Martin Reuter, Franz-Erich Wolter, Niklas Peinecke, Laplace–Beltrami spectra as 'shape-DNA'of surfaces and solids, Comput.-Aided Des. 38 (4) (2006) 342–366.

[17] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, Olaf Ronneberger, 3D U-net: learning dense volumetric segmentation from sparse annotation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, Springer, 2016, pp. 424–432.

[18] Andriy Myronenko, 3D MRI brain tumor segmentation using autoencoder regularization, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4, Springer, 2019, pp. 311–320.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[20] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, Daniel Rueckert, Attention gated networks: Learning to leverage salient regions in medical images, Med. Image Anal. 53 (2019) 197–207.

[21] Sigurd Vangen Wifstad, Henrik Agerup Kildahl, Bjørnar Grenne, Espen Holte, Ståle Wågen Hauge, Sigbjørn Sæbø, Desalew Mekonnen, Berhanu Nega, Rune Haaverstad, Mette-Elise Estensen, et al., Mitral valve segmentation and tracking from transthoracic echocardiography using deep learning, Ultrasound Med. Biol. (2024).

[22] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, Daguang Xu, Unetr: Transformers for 3d medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 574–584.

[23] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, Kilian Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[24] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, Leonardo Rundo, Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation, Comput. Med. Imaging Graph. 95 (2022) 102026.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, Soumith Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d´Alché-Buc, E. Fox, R. Garnett (Eds.), in: Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019, pp. 8024–8035.

[26] M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al., MONAI: An open-source framework for deep learning in healthcare, 2022, arXiv preprint arXiv:2211.02701.

[27] Zhengyou Zhang, Iterative point matching for registration of free-form curves and surfaces, Int. J. Comput. Vis. 13 (2) (1994) 119–152.

[28] Will Schroeder, Ken Martin, Bill Lorensen, The Visualization Toolkit, fourth ed., Kitware, ISBN: 978-1-930934-19-1, 2006.

[29] Fei Jia, Shu Wang, V.T. Pham, A hybrid catheter localisation framework in echocardiography based on electromagnetic tracking and deep learning segmentation, Comput. Intell. Neurosci. 2022 (1) (2022) 2119070.

[30] Riccardo Munafò, Simone Saitta, Davide Tondi, Giacomo Ingallina, Paolo Denti, Francesco Maisano, Emiliano Votta, et al., Automatic 4D mitral valve segmentation from transesophageal echocardiography: a semi-supervised learning approach, Med. Biol. Eng. Comput. (2025) 1–16.

[31] A. Vaswani, Attention is all you need, Adv. Neural Inf. Process. Syst. (2017).

[32] Santiago Garcia, Sammy Elmariah, Robert J. Cubeddu, Firas Zahr, Mackram F. Eleid, Susheel K. Kodali, Puvi Seshiah, Rahul Sharma, D. Scott Lim, Mitral transcatheter edge-to-edge repair with the PASCAL precision system: Device knobology and review of advanced steering maneuvers, Struct. Hear. 8 (1) (2024) 100234.

[33] Paul Sorajja, Brian Whisenant, Nadira Hamid, Hursh Naik, Raj Makkar, Peter Tadros, Matthew J. Price, Gagan Singh, Neil Fam, Saibal Kar, et al., Transcatheter repair for patients with tricuspid regurgitation, N. Engl. J. Med. 388 (20) (2023) 1833–1842.

[34] Charles Ruizhongtai Qi, Li Yi, Hao Su, Leonidas J. Guibas, Pointnet++: Deep hierarchical feature learning on point sets in a metric space, Adv. Neural Inf. Process. Syst. 30 (2017).