# Model-to-Image Registration via Deep Learning towards Image-Guided Endovascular Interventions

Zhen Li[1,2], Maria Elisabetta Mancini[3], Giovanni Monizzi[3], Daniele Andreini[3,4], Giancarlo Ferrigno[1], Jenny Dankelman[2] and Elena De Momi[1]

2021 International Symposium on Medical Robotics (ISMR) | 978-1-6654-0622-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISMR48346.2021.9661511

***Abstract*— Cardiologists highlight the need for an intra-operative 3D visualization to assist interventions. The intra-operative 2D X-ray/Digital Subtraction Angiography (DSA) images in the standard clinical workflow limit cardiologists' views significantly. Compared with image-to-image registration, model-to-image registration is an essential approach taking advantage of the reuse of pre-operative 3D models reconstructed from Computed Tomography Angiography (CTA) images. Traditional optimized-based registration methods suffer severely from high computational complexity. Moreover, the consequence of lacking ground truth for learning-based registration approaches should not be neglected. To overcome these challenges, we introduce a model-to-image registration framework via deep learning for image-guided endovascular catheterization. This work performs autonomous vessel segmentation from intra-operative fluoroscopy images via a deep residual U-net and a model-to-image matching via a convolutional neural network. For this study, image data were collected from 10 patients who performed Transcatheter Aortic Valve Implantation (TAVI) procedures. It was found that vessel segmentation of test data results in median values of Dice Similarity Coefficient, Precision, and Recall of (0.75, 0.58, 0.67) for femoral artery, and (0.71, 0.56, 0.74) for aortic root. The segmentation network behaves better than manual annotation, and it recognizes part of vessels that were not labeled manually. Image matching between the transformed moving image and the fixed image results in a median value of Recall of 0.90. The proposed approach achieves a good accuracy of vessel segmentation and a good recall value of model-to-image matching.**

***Index Terms*— Endovascular Catheterization, Image-guided Interventions, Image Registration, Deep Learning, Convolutional Neural Network**

## I. INTRODUCTION

Aortic valve stenosis is a narrowing of the aortic valve opening. It is a common, abnormal condition of heart's aortic valve, and it can be severe. Narrowed valve, typically due to calcium deposits, will affect blood circulation [1]. Transcatheter Aortic Valve Implantation (TAVI) is a procedure for treating patients with aortic valve stenosis by placing a trans-catheter prosthetic valve at the aortic valve. The transfemoral access approach is the choice in the vast majority of TAVI patients (around 70%) [2].

Intra-operatively, the access route and target site are essential imaging regions to guarantee the safety of needle insertion and accuracy of prosthetic valve alignment. Fluoroscopy imaging is the traditional intra-operative imaging approach. Nevertheless, that 2D image does not provide enough information for cardiologists, and the need for an intra-operative 3D visualization is highlighted. For example, the clinical study in [3] stressed that image fusion could help cardiologists for performing TAVI procedures and validated it through controlled experiments. The fusion image of the aortic root is presented intra-operatively under the support of the *Valve ASSIST 2* (GE Healthcare), which is registered at the beginning of the procedure from multiple fluoroscopy views. In addition to facilitating the operation of cardiologists, image registration between various imaging modality methods can also significantly improve the guidance of steerable instruments (for example, magnetically actuated robotic catheters) [4]. The image registration can also help for autonomous target localization for robotic interventions and increase the level of autonomy [5].

The state-of-the-art image registration is classified into optimization-based and learning-based approaches. For optimization-based approaches, the aortic centerline [6] and aortic shape contour [7] are commonly used as features for matching. In [6], a graph matching method is proposed to establish the correspondence between the 3D pre-operative and 2D intra-operative skeletons extracting from fluoroscopic images, and then the two skeletons are registered by skeleton deformation. The work in [7] estimated a warp field of 3D aortic shape deformation by solving a non-linear least-squares problem based on an embedded deformation graph. However, the optimization-based approaches suffer severely from high computational complexity, and learning-based approaches are explored and reviewed in [8]. A model-image registration of left ventricle by imitation learning is proposed in [9] for cardiac resynchronization therapy. In the reported results of [9], the target registration error was measured by computing the L2 norm of the points of the cross landmark at the center, between the fluoroscopy cross and the registered left ventricle model cross, and an error of $2.92 \pm 2.22$ mm was reported. A model-image registration via multi-channel Convolutional Neural Network (CNN) with deformations caused by heartbeat and respiration is introduced by [10], where different channels represent different phases in the diameter periodic variation. Nevertheless, it needs a dataset

*This work was supported by the ATLAS project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 813782.

[1]NearLab, Department of Electronics, Information and Bioengineering(DEIB), Politecnico di Milano, Milan, Italy zhen.li@polimi.it

[2]Department of Biomechanical Engineering, Delft University of Technology, Delft, The Netherlands

[3]Centro Cardiologico Monzino, IRCCS, Milan, Italy

[4]Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy

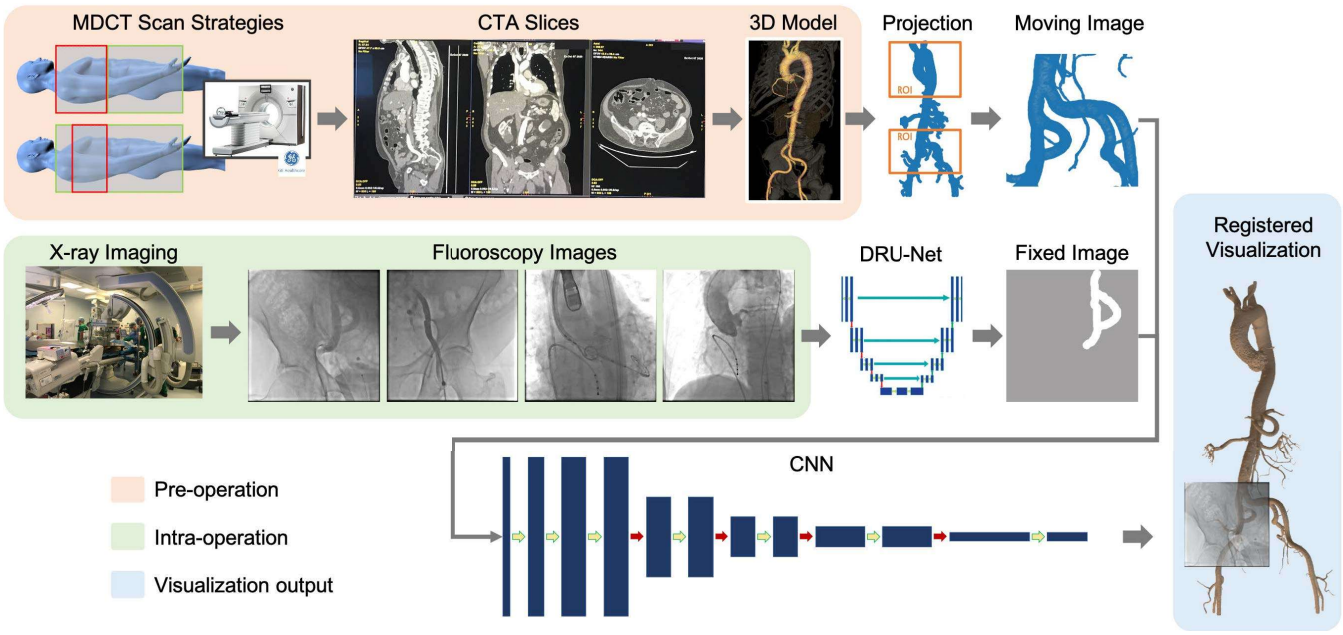


Fig. 1. The workflow of model-to-image registration via deep learning. Pre-operatively, a 3D model is reconstructed from CTA slices, and a projection image is generated afterward. Autonomous vessel segmentation from intra-operative fluoroscopy images is implemented via DRU-Net, and model-to-image matching via CNN provides a registered visualization.

including complete cycles for learning the change of vessel diameter, which entails a long time exposure for imaging and a great amount of contrast media used.

The research described in literature using optimization-based approaches suffered severely from high computational complexity [6]. Most studies on learning-based approaches in literature use the landmarks or other annotated features as ground truth, which takes a tremendous effort for large datasets [8]. The contributions of this work are summarized as follows. First, we present a novel framework using deep learning for image registration based on the pre-operative model instead of image slices. Second, we propose a novel model-to-image matching approach in an unsupervised way, i.e., no landmarks or other annotated features for matching are provided.

This paper introduces a novel model-to-image registration framework via deep learning for image-guided endovascular catheterization. In order to find the correspondence between a pre-operative model and intra-operative images, this framework firstly performs autonomous vessel segmentation from fluoroscopy images. A deep residual U-net architecture is employed, thanks to its fast learning convergence and efficient spatial information propagation without degradation. After that, a model-to-image matching via CNN is introduced. The proposed framework is based on the reuse of the pre-operative model that is reconstructed from Computed Tomography Angiography (CTA) images for diagnosing and size measurement. For this study, image data were collected from 10 patients who performed TAVI procedures, and in total, 1221 annotated fluoroscopy frames were used.

The rest of this paper is structured as follows. The materials and model-to-image registration approach are introduced in Section II. Section III discusses the experimental results. Conclusions and future work are presented in Section IV.

## II. MATERIALS AND METHODOLOGY

The framework of the proposed model-to-image registration approach is introduced in Fig. 1. It presents the workflow to obtain a pre-operative 3D model (detailed in Sec. IIA), autonomous vessel segmentation from intra-operative fluoroscopy images (detailed in Sec. IIB), and model-to-image matching (detailed in Sec. IIC).

For this study, image data were collected from 10 patients who performed TAVI procedures. The data were acquired from the Centro Cardiologico Monzino (CCM) at Milan, Italy. The data collection followed the ethical protocol approved by the CCM under the assigned code of *02_21 PA*.

### A. 3D Model Reconstruction

The pre-operative CTA images were acquired following two typical Multidetector Computed Tomography (MDCT) scan strategies: cardiac Electrocardiogram (ECG)-syncronized CTA of the aortic root and heart followed by a non-ECG-syncronized helical CTA of the thorax, abdomen and pelvis; ECG-syncronized CTA of the thorax followed by a non-ECG-syncronized helical CTA of the abdomen and pelvis. Respiratory motion is also a common artifact seen at cardiac CT [11], and there are novel studies regarding motion correction under a free-breathing acquisition mode [12], [13]. In this study, a breath-holding method was employed for CT scan acquisition, and the respiratory motion was negligible theoretically.

Semi-automatic segmentation of the vessels and 3D mesh model reconstruction were performed using the *AW server* (GE Healthcare), followed by a manual refinement process.

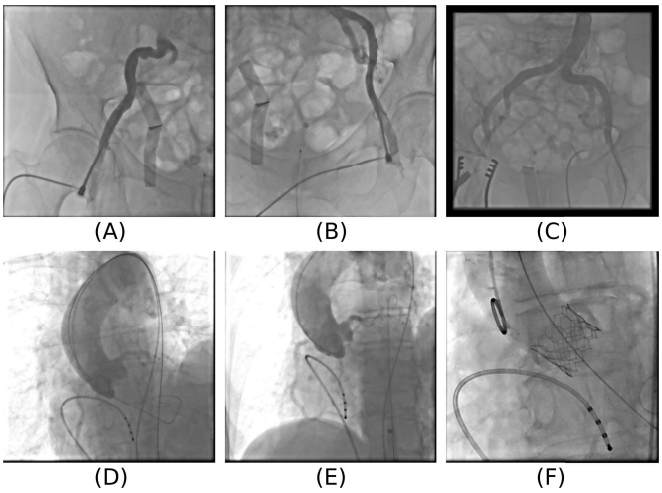


Fig. 2. Sample of fluoroscopy images in the dataset collected showing the variability of the FoV: (A) right femoral, (B) left femoral, (C) aortic bifurcation, and (D-F) aortic root.

TABLE I

DETAILED INFORMATION ABOUT THE DATASET COLLECTED.

| Patient No. | lower FoV | | upper FoV | |
|---|---|---|---|---|
| | No. of annotated frames | Image Size (pixels) | No. of annotated frames | Image Size (pixels) |
| 1 | 52 | 512*512 | 47 | 512*512 |
| 2 | 37 | 512*512 | 62 | 512*512 |
| 3 | 43* | 512*512 | 124 | 512*512 |
| 4 | 20 | 512*512 | 23 | 512*512 |
| 5 | 18 | 512*512 | 45 | 512*512 |
| 6 | 20 | 1024*1024 | 101* | 512*512 |
| 7 | 57 | 512*512 | 60 | 512*512 |
| 8 | 48 | 512*512 | 178 | 512*512 |
| 9 | 70 | 512*512 | 81 | 512*512 |
| 10 | 38 | 512*512 | 97 | 512*512 |
| Training | 275 | — | 569 | — |
| Validation | 85 | — | 148 | — |
| Testing | 43 | — | 101 | — |
| Total | 403 | — | 818 | — |

*These images were set apart to be used only during testing.

The 3D models were exported under the support of the *3D suite* (GE Healthcare).

### B. 2D Shape Segmentation

A large number of methods have been proposed to automatically or semi-automatically segment vascular structures [14]. The residual networks have been proven to achieve better performance than state-of-the-art algorithms for vessel segmentation tasks [15]. In this work, we employ a Deep Residual U-Net (DRU-Net) architecture [16] for vessel segmentation in fluoroscopy images.

The fluoroscopy images were intra-operatively acquired. The Field-of-View (FoV) was changed during the procedure to show the concerned vessels condition. Some samples of the dataset images which depict the variability of the FoV are shown in Fig. 2. As a summary, the images can be classified into two categories: the frames with a lower FoV (see Fig. 2A-C), and the frames with an upper FoV (see Fig. 2D-F). From original fluoroscopy images, a different number of frames were extracted as described in TABLE I and manually annotated using *PixelAnnotationTool* [17].

A prepossessing stage included resizing of the frames to 256×256 to be consistent with the input layer of the DRU-Net used. This training was done by using all the annotated frames obtained from 9 patients for training and validation, while the data from one patient were used for testing. The training, validation, and testing dataset in this stage were arranged as illustrated in TABLE I.

The DRU-Net segmentation network adapts a 2D encoder-decoder architecture taking an input layer size of 256×256. The encoder consists of four residual blocks, each followed by a 2×2 max-pooling layer. Each residual block uses two convolutional layers of kernel size 3×3, each followed by batch normalization and a ReLU activation function. A residual connection is introduced for each residual block, which adds the input to the output of the same block. The decoder consists of up-sampling of feature maps from the lower level and concatenation with the feature maps from the corresponding contracting path.

For the training on frames with a lower FoV, the loss function is based on the Dice Similarity Coefficient (DSC),

$$\mathcal{L}_{(segl)} = 1 - \frac{2TP}{2TP + FN + FP} \quad (1)$$

where $TP$ is the number of pixels that belong to the vessels, which are corrected segmented, $FP$ is the number of pixels miss-classified as vessels and $FN$ is the number of pixels that should be classified as vessels, but actually, they are not.

For the training on frames with an upper FoV, the loss function is based on a binary crossentropy defined as:

$$\mathcal{L}_{(segu)} = -\frac{1}{N}\sum_{i=1}^{N} Y_i \log Y_i' + (1 - Y_i)\log(1 - Y_i') \quad (2)$$

where $N$ is the number of scalar values in the model output, $Y_i'$ is the $i$th scalar value in the model output and $Y_i$ is the corresponding target value.

### C. Model-to-Image Matching

Given a pre-operative 3D model, a projection image is generated according to the primary and secondary angles

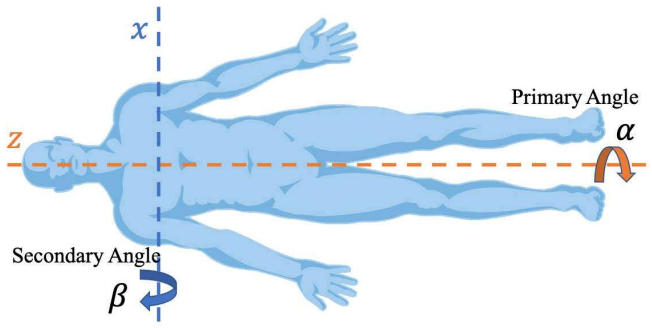


Fig. 3. Scheme of the primary and secondary angles in the fluoroscopy metadata.

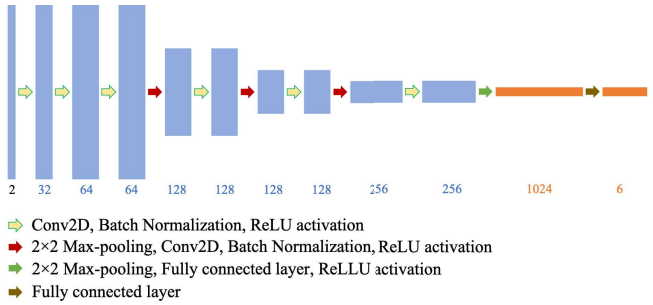


Fig. 4. The sketch of the CNN architecture for model-to-image matching.

obtained from the fluoroscopy metadata. The definition of these two angles is shown in Fig. 3. Thus the projection can be expressed as:

$$\begin{bmatrix} x_p \\ y_p \\ d_p \end{bmatrix} = \mathrm{Rot}_z(\alpha)\mathrm{Rot}_x(\beta) \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{3}$$

where $\alpha$ and $\beta$ are the primary angle and secondary angle, respectively. $\mathrm{Rot}_z$, $\mathrm{Rot}_x$ denotes the rotation matrix. $(x, y, z)$ is the coordinates of mesh model vertices and $(x_p, y_p)$ is the coordinates of projected vertices.

From a projection image, a Region of Interest (ROI) is interactively selected and referred to as the moving image for further image matching.

With deep learning, given a pair of moving and fixed images, the registration network outputs an affine transformation matrix $\boldsymbol{M}$, which can be considered as a combination of scaling, shearing, translation, rotation of the moving image. That matrix defines a mapping from the moving image's coordinates to that of the fixed image, which can be expressed as:

$$\begin{bmatrix} x_f \\ y_f \\ 1 \end{bmatrix} = \boldsymbol{M} \begin{bmatrix} x_m \\ y_m \\ 1 \end{bmatrix} \text{ with } \boldsymbol{M} = \begin{bmatrix} m_{11} & m_{12} & d_x \\ m_{21} & m_{22} & d_y \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

where $(x_f, y_f)$ is the coordinates of the fixed image, $(x_m, y_m)$ is that of the moving image and $\boldsymbol{M}$ is the transformation matrix.

The CNN adopts a 2D encoder architecture as shown in Fig. 4, taking an input layer size of $256 \times 256$. The convolutional layers use 32, 64, 64, 128, 128, 128, 128, 256, 256 filters. The kernel size of the first convolutional layer is $5 \times 5$, and it is $3 \times 3$ for the others. Each convolutional layer is followed by batch normalization and a ReLU activation function. From the third convolutional layer, it is followed by a $2 \times 2$ max-pooling layer every other time. The first fully connected layer has 1024 neurons with a ReLU activation function. The last fully connected layer has six neurons representing transformation matrix parameters.

For training the network without the ground truth of the transformation matrix, the moving image is transformed respecting the predicted transformation matrix output, and the difference between the transformed image and the fixed image is regarded as the training loss. A bilinear interpolation layer from Spatial Transform Network (STN) [18] is implemented for performing that transform. The loss function is therefore defined with respect to the DSC between the fixed image and the predicted transformed image.

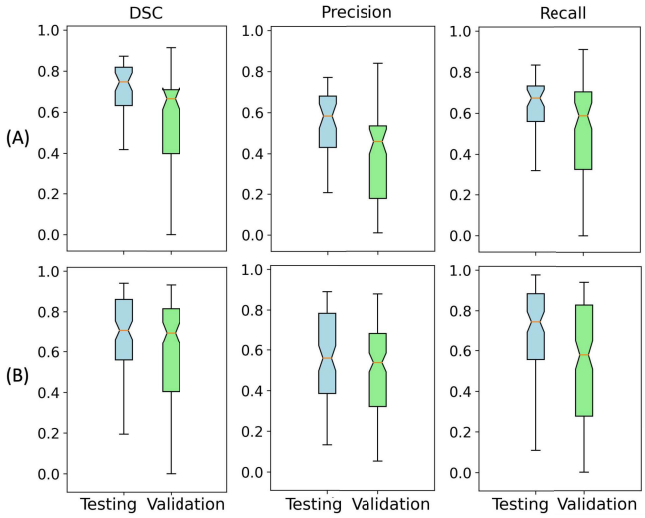


Fig. 5. Box plots obtained from the segmentation results of frames (A) with a lower FoV and (B) with an upper FoV, with respect to DSC, Precision, and Recall.

$$\mathcal{L}_{(reg)} = 1 - \frac{2TP}{2TP + FN + FP} \tag{5}$$

where $TP$ is the number of pixels that belong to the vessels in the fixed image, which are corrected registered, $FP$ is the number of pixels miss-registered as vessels, and $FN$ is the number of pixels that should be registered as vessels, but actually, they are not.

### D. Experiment and Validation

The performance metrics chosen for evaluating the segmentation and registration results were the DSC, the Precision and Recall which are defined as

$$DSC = \frac{2TP}{2TP + FN + FP} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

The learning rate and mini-batch size of the segmentation network DRU-Net were chosen by trying the different combinations between several possible values of the hyper-parameters. The training on images with a lower FoV used a learning rate of $1e{-}5$ and a batch size of 8. The training on images with an upper FoV used a learning rate of 0.1 and a batch size of 4. The learning rate and mini-batch size of the model-to-image matching network CNN were $1e{-}4$ and 4, respectively.

The Networks were implemented using *Tensorflow* and *Keras* frameworks in *Python* trained on a NVIDIA GeForce RTX2080Ti GPU card.

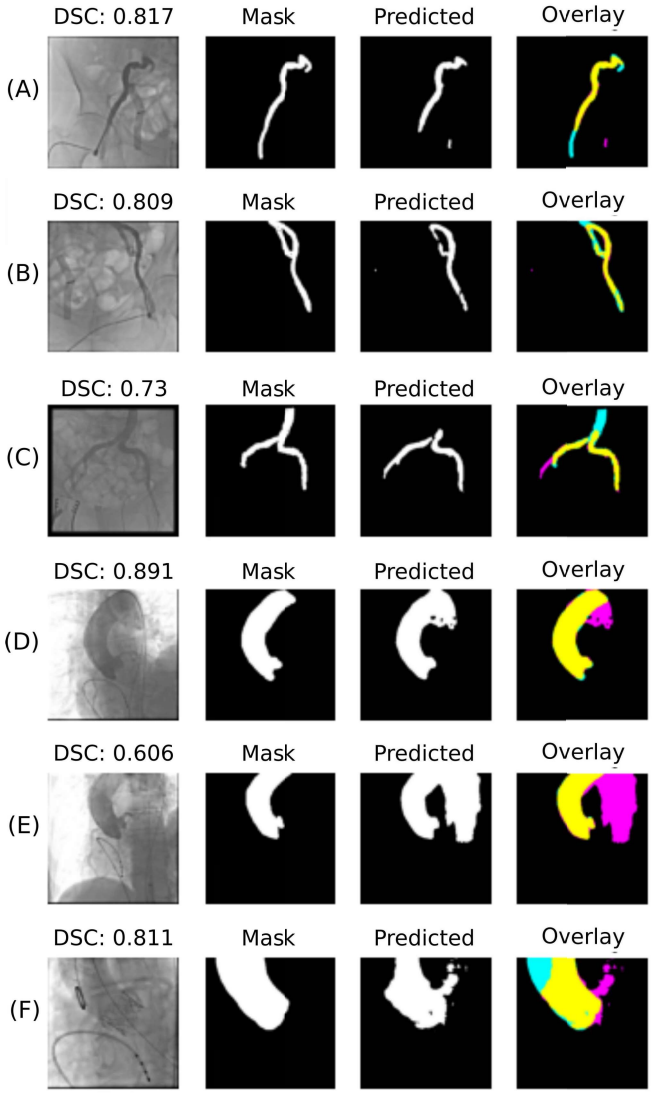


Fig. 6. Sample of results obtained using DRU-Net with its respective dice value. The colors in the Overlay images are as follows. $TP$: Yellow, $FP$: Magenta, $FN$: Cyan

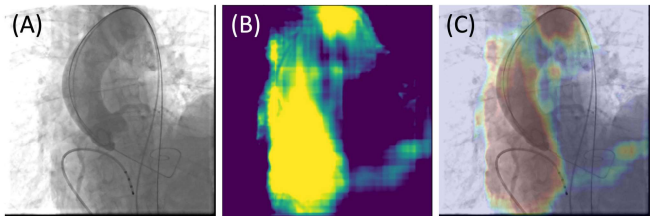


Fig. 7. The resulting Grad-CAM attention map from the input image, which produces a coarse localization map highlighting the important regions in the image for predicting the segmentation. (A) Input image (B) Grad-CAM attention map (C) Overlapping image between (A) and (B).

## III. RESULTS AND DISCUSSION

### A. 2D shape Segmentation

The box plots obtained from testing and validation data with respect to DSC, Precision and Recall are shown in Fig. 5. It was found that vessel segmentation of test data with a lower FoV results in median values of DSC, Precision, and Recall of 0.75, 0.58, and 0.67, respectively. The segmentation of test data with an upper FoV gives good results obtaining median values of DSC, Precision, and Recall of 0.71, 0.56, and 0.74, respectively.

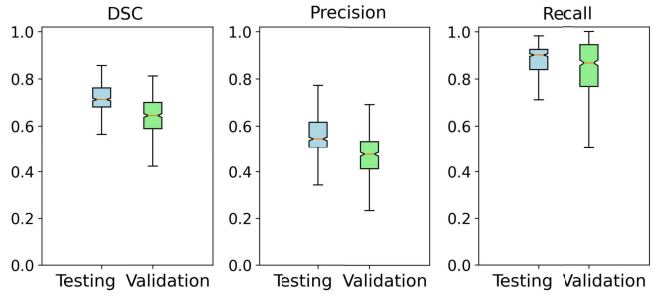


Fig. 8. Box plots obtained from the results of image matching with respect to DSC, Precision, and Recall.

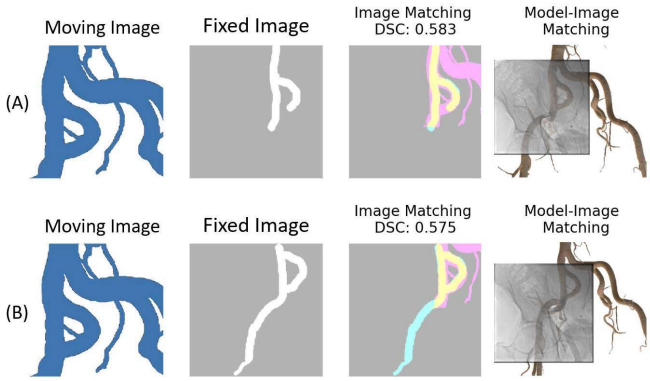


Fig. 9. Sample of results obtained using CNN for model-to-image registration. The colors in the image matching results are as follows. $TP$: Yellow, $FP$: Magenta, $FN$: Cyan

Some samples of the results obtained with DRU-Net are shown in Fig. 6. It shows that some parts of the vessels are correctly segmented in the majority of the cases. Fig. 6A shows that sometimes the network can not distinguish vessels and devices (tube structure). In some cases (Fig. 6D-F), the network behaves better than manual annotation (i.e., mask), and it recognizes parts of vessels that were not labeled manually.

A Grad-CAM attention map [19] was generated to produce 'visual explanations' for decisions from the DRU-Net model and to make them more transparent. The Grad-Cam approach creates a coarse localization map highlighting the critical regions in the image for predicting the segmentation. For example in Fig. 7, the aortic arch is detected as an important region, while the coronaries regions are not considered as critical regions in this case.

### B. Model-to-Image Registration

The box plots obtained from testing and validation data with respect to DSC, Precision and Recall are shown in Fig. 8. Image matching between the transformed moving image and the fixed image results in median values of DSC, Precision, and Recall of 0.71, 0.55, and 0.90, respectively.

Since the fixed image might cover only partial branches of the vasculature, the Recall value can better reveal matching accuracy. Some samples of the results obtained with CNN are shown in Fig. 9. The model-to-image registration can provide a 3D visualization to guide and locate the device more visually.

The deformation field is neglected in this work, considering that the autonomous vessel segmentation is not always accurate. The lack of segmentation accuracy would affect the deformation detection significantly. Therefore the model with deformation would be less stable and could increase the workload and disturbance to cardiologists.

In our datasets, the subjects took the procedure in 0-7 months (some delayed due to COVID19) after the pre-operative CTA scanning. The proposed framework is based on the reuse of the pre-operative model that is reconstructed from CTA images for diagnosing and size measurement. We assume that this pre-operative model is still accurate at the moment of the actual procedure date.

There is a factor that might influence the performance of this approach: the accuracy of the reconstructed model. The pre-operative imaging modality does not greatly influence the registration performance, but the reconstructed model could have an influence.

## IV. CONCLUSIONS

This work proposed a model-to-image registration framework via deep learning for image-guided endovascular interventions. Autonomous vessel segmentation from intra-operative fluoroscopy images is implemented via DRU-Net, and model-to-image matching via CNN provides a registered visualization. The proposed framework is based on the reuse of the pre-operative model reconstructed from CTA images for diagnosing and size measurement, and there is no interference with the standard clinical workflow. For this study, image data were collected from 10 patients who performed TAVI procedures, and in total, 1221 annotated fluoroscopy frames were used. The results show that the proposed approach achieves a good accuracy of vessel segmentation and a good recall value of model-to-image matching.

Future work improvements include integrating augmented reality, performing user-end evaluation in the operating room, and extending to a deformable registration approach considering the vessel deformations due to the device contact during the procedure. Moreover, the authors would improve the datasets by providing the landmarks or other annotated features as ground truth for image matching, and then compare the performance of the proposed approach with the works of literature.

## ACKNOWLEDGMENT

This work was supported by the ATLAS project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 813782. We gratefully acknowledge the support of the Centro Cardiologico Monzino with the resources of the clinical data used for this research.

## REFERENCES

[1] V. Mallika, B. Goswami, and M. Rajappa, "Atherosclerosis pathophysiology and the role of novel risk factors: A clinicobiochemical perspective," *Angiology*, vol. 58, no. 5, pp. 513–522, 2007.

[2] L. Biasco, E. Ferrari, G. Pedrazzini, F. Faletra, T. Moccetti, F. Petracca, and M. Moccetti, "Access sites for tavi: patient selection criteria, technical aspects, and outcomes," *Frontiers in cardiovascular medicine*, vol. 5, p. 88, 2018.

[3] C. Butter, H. Kaneko, G. Tambor, M. Hara, M. Neuss, and F. Hoelschermann, "Clinical utility of intraprocedural three-dimensional integrated image guided transcatheter aortic valve implantation using novel automated computed tomography software: A single-center preliminary experience," *Catheterization and Cardiovascular Interventions*, vol. 93, no. 4, pp. 722–728, 2019.

[4] C. Heunis, J. Sikorski, and S. Misra, "Flexible instruments for endovascular interventions: improved magnetic steering, actuation, and image-guided surgical instruments," *IEEE robotics & automation magazine*, vol. 25, no. 3, pp. 71–82, 2018.

[5] A. Attanasio, B. Scaglioni, E. De Momi, P. Fiorini, and P. Valdastri, "Autonomy in surgical robotics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, pp. 651–679, 2021.

[6] J.-Q. Zheng, X.-Y. Zhou, C. Riga, and G.-Z. Yang, "Towards 3d path planning from a single 2d fluoroscopic image for robot assisted fenestrated endovascular aortic repair," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8747–8753.

[7] Y. Zhang, R. Falque, L. Zhao, S. Huang, and B. Hu, "Deep learning assisted automatic intra-operative 3d aortic deformation reconstruction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 660–669.

[8] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: a survey," *Machine Vision and Applications*, vol. 31, no. 1, pp. 1–18, 2020.

[9] D. Toth, S. Miao, T. Kurzendorfer, C. A. Rinaldi, R. Liao, T. Mansi, K. Rhode, and P. Mountney, "3d/2d model-to-image registration by imitation learning for cardiac procedures," *International journal of computer assisted radiology and surgery*, vol. 13, no. 8, pp. 1141–1149, 2018.

[10] S. Guan, C. Meng, Y. Xie, Q. Wang, K. Sun, and T. Wang, "Deformable cardiovascular image registration via multi-channel convolutional neural network," *IEEE Access*, vol. 7, pp. 17 524–17 534, 2019.

[11] K. Kalisz, J. Buethe, S. S. Saboo, S. Abbara, S. Halliburton, and P. Rajiah, "Artifacts at cardiac ct: physics and solutions," *Radiographics*, vol. 36, no. 7, pp. 2064–2083, 2016.

[12] Z. Liu, Z. Zhang, N. Hong, L. Chen, C. Cao, J. Liu, and Y. Sun, "Diagnostic performance of free-breathing coronary computed tomography angiography without heart rate control using 16-cm z-coverage ct with motion-correction algorithm," *Journal of cardiovascular computed tomography*, vol. 13, no. 2, pp. 113–117, 2019.

[13] J. Liang, Y. Sun, Z. Ye, Y. Sun, L. Xu, Z. Zhou, B. Thomsen, J. Li, Z. Sun, and Z. Fan, "Second-generation motion correction algorithm improves diagnostic accuracy of single-beat coronary ct angiography in patients with increased heart rate," *European radiology*, vol. 29, no. 8, pp. 4215–4227, 2019.

[14] S. Moccia, E. De Momi, S. El Hadji, and L. S. Mattos, "Blood vessel segmentation algorithms—review of methods, datasets and evaluation metrics," *Computer methods and programs in biomedicine*, vol. 158, pp. 71–91, 2018.

[15] D. Li, D. A. Dharmawan, B. P. Ng, and S. Rahardja, "Residual u-net for retinal vessel segmentation," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1425–1429.

[16] J. F. Lazo, A. Marzullo, S. Moccia, M. Catellani, B. Rosa, M. de Mathelin, and E. De Momi, "A lumen segmentation method in ureteroscopy images based on a deep residual u-net architecture," *Proceedings of the 26th International Conference on Pattern Recognition (ICPR)*, 2021.

[17] A. Bréhéret, "Pixel Annotation Tool," https://github.com/abreheret/PixelAnnotationTool, 2017.

[18] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *arXiv preprint arXiv:1506.02025*, 2015.

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.