A Multi-Task Deep Neural Network for Segmentation and Landmark Detection in Cardiac Computerized Tomography

Nicla Mandas[1,2], Giulia Baldazzi[2], Andrea Pitzus[2], Giacomo Tarroni[3,4], Danilo Pani[2]

[1]Hadron Academy, Istituto Universitario di Studi Superiori, IUSS, Pavia, Italy
[2]Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy
[3]CitAI Research Centre, Department of Computer Science, City St George's, University of London, London, UK
[4]BioMedIA, Department of Computing, Imperial College London, London, UK

Abstract

Multimodal bioimaging is increasingly recognized for its potential to integrate multiple types of information. This is particularly relevant in interventional cardiology, where structural imaging may be fused with complementary data, such as metabolic or electrophysiological data. Automating the preprocessing steps required for image alignment and registration is crucial to accelerate procedures in clinical settings.

This study explores the feasibility of using a multi-task deep neural network for the automatic segmentation of the left ventricle from cardiac computerized tomography scans and the prediction of a landmark position required for image alignment. The model, based on a 3D UNet architecture, simultaneously segments the left ventricle and localizes its apex. It was trained and tested on the segmented images of the Multi-Modality Whole Heart Segmentation dataset, where the apex position was manually annotated by an expert.

The network achieved an average Dice score of 0.91 and an average Euclidean distance of 11 mm for the segmentation and the landmark detection, respectively.

1. Introduction

Cardiac imaging plays a crucial role in the diagnosis and treatment of cardiovascular diseases. However, different imaging techniques may provide different insights into cardiac anatomy and function. Indeed, structural imaging, such as computed tomography (CT) or cardiac magnetic resonance (CMR), provides detailed anatomical information, but can fail in fully capturing functional information. On the other hand, complementary functional techniques, like the electroanatomic (EA) mapping, although supporting cardiac electrophysiological procedures, provide less accurate anatomical information. Hence, integrating multiple information into a single, multimodal image is often desired [1]. Hybrid imaging, e.g., PET/CT or SPECT/CT, has improved diagnostic robustness in coronary artery disease by combining perfusion data with coronary anatomy [2]. Similarly, the fusion of CMR-derived scar maps with EA maps has refined ablation strategies for ventricular tachycardia [3].

In this scenario, deep learning methods may offer a paradigm shift, enabling end-to-end architectures to address multiple tasks like image segmentation and landmark detection concurrently. Notably, 3D UNet-based models represent one of the best choices for automatic cardiac segmentation [4]; however, the integration of auxiliary tasks (e.g., landmark prediction) remains unexplored.

In light of these premises, in this work we propose a 3D multi-task deep neural network for the simultaneous segmentation of the left ventricle and the detection of its apex in cardiac CT scans. The model, built on a 3D UNet architecture, is assessed on a public dataset, proving its potential as an automated preprocessing step for multimodal pipelines.

2. Methods

2.1. Dataset

The Multi-Modality Whole Heart Segmentation dataset [5], [6], [7], [8] was used in this study. It comprises anonymized clinical magnetic resonance imaging (MRI) and CT scans for whole heart segmentation, which were performed in-vivo during routine clinical procedures. Thus, image quality was not uniform across the dataset.

CT scans were acquired during routine cardiac CT angiography, covering the whole heart from the upper abdomen to the aortic arch, with slices acquired in the axial view. The dataset includes 20 labeled images, originally conceived for the training set, and 40

unlabelled images, in turn conceived for the test set. Given the purpose of this study, only the labeled images were considered, and we focused on the left ventricle blood cavity for the segmentation. The in-plane resolution was on average 0.429×0.429 mm, while the slice thickness was either 0.625 mm (fifteen images) or 0.45 mm (five images). Image size was 512×512 in the 2D plane, with the number of slices varying from 177 to 363.

To pursue the research goal, we added to the segmentation the annotation of a landmark indicating the apex of the left ventricle. This landmark was marked by an expert on CT scans via the ITK-SNAP application [9], by applying a small sphere centered on the selected pixel.

2.2. Data preprocessing and augmentation

A set of different transformations was initially applied. After loading the CT scans and assuring that they were in the same format, the left ventricle's mask was extracted from the provided segmentation. The landmark coordinates were obtained by computing the sphere center added with ITK-SNAP onto the mask. Then, a heatmap was obtained by computing the Euclidean distance, with zeros representing the landmark location, followed by a logarithmic scaling of the values to highlight the landmark position. Finally, heatmaps were rescaled in the range [0,1] by applying the min-max normalization.

Data augmentation was implemented on the training set via random cropping. Sub-volumes of size 128×128×128 were extracted while ensuring a balance between positive and negative samples, where the former were regions containing the target label (i.e., the left ventricle), while the latter represented background or non-target regions.

2.3. Network architecture

The proposed deep learning model is a multi-task network based on a 3D UNet architecture implemented within the MONAI framework, which includes both segmentation of the left ventricle and landmark localization, inspired by the work reported in [10]. Its architecture is represented in Fig. 1.

The network takes as input the CT scan as a 3D tensor, which is then fed into the encoding section, composed of five levels for feature extraction and landmark detection.

Each level includes an encoding block formed by two residual units and a downsampling stage, performed with a strided convolution with a stride of 2, which takes the initial feature maps from 32 up to 512. In the decoding section, there are upsampling stages performed with a transpose convolution, always with a stride of 2, followed by two residual units (i.e., the decoding block). Finally, a last convolutional layer is applied to reduce the number of channels. Two different activation functions are then employed for the two tasks: a softmax function for the segmentation of the left ventricle, and a sigmoid function for the landmark heatmap. To enhance model generalization, dropout regularization is incorporated with a probability of 0.1. The skip connections between the encoder and the decoder parts, indicated by the dashed lines in Fig. 1, preserve spatial information, for accurate segmentation and landmark localization.

2.4. Training and evaluation strategy

A 10-fold cross-validation was performed. Specifically, for each fold, labeled images in the dataset were partitioned into training, validation, and test sets following an 80/10/10 split. We ensured that, in each fold, the testing subjects couldn't also be present in the

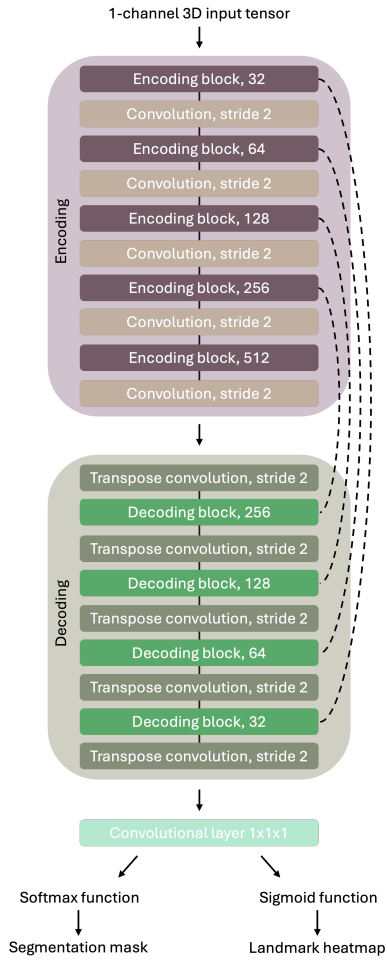

Figure 1. Model architecture, based on a 3D UNet. Input is represented at the top, followed by the encoding block (light purple), the decoding block (light green), the last convolutional layer, and the two outputs of the network. Skip connections between the encoding and decoding blocks are represented with the dashed lines.

training and/or validation sets.

After several preliminary tests aimed at evaluating the behaviour of the losses during the training process, it was decided that the implemented UNet would be trained for 1000 epochs in each fold. The loss for the segmentation task was a weighted sum of the Dice loss and the Cross Entropy loss (DiceCELoss), while for the landmark prediction, the Mean Squared Error Loss (MSELoss) was chosen. The total loss used for backpropagation was computed as the weighted sum of the individual losses for the two tasks, with the weight being $\alpha=1$ for the segmentation task, and $\beta=10$ for the landmark prediction. Validation was performed every other epoch, during which the model was saved every time it reached a new minimum value in the total loss.

A sliding window inference was applied in the validation loop, with an overlap between the sub-volumes of 25%, to revert to the original image size, given that the network was exposed to sub-volumes (i.e., patches) during training rather than the entire image. The network gave a heatmap as output for the landmark prediction, from which the landmark's coordinates in pixels were extracted by taking the coordinates of its minimum value.

To evaluate the performance of the network, two metrics have been adopted: the Dice score for the segmentation task and the Euclidean distance (i.e., L2 distance) for the landmark coordinates. Given the different resolutions of the images, we multiplied every coordinate for the corresponding resolution before computing the Euclidean distance.

Data analysis was performed with Python v3.9.16, using the Microsoft Visual Studio Code IDE on a high-performance computing (HPC) cluster, with a computational node made of 4 NVIDIA A100 GPUs. Each GPU is configured with 6912 CUDA cores and 80 GB of high-bandwidth memory, providing enough space for the computational workload. PyTorch and MONAI were used for the deep learning model implementation and the processing of medical images, respectively.

3. Results and discussion

In Table 1, the training duration, the epoch at which the best model was saved, and the corresponding loss are reported for each fold. Overall, the training time was 16342±466 s, and the lowest loss reached 0.087±0.025, with substantial consistency across folds.

In Table 2, the performance of the network on the images of the test sets is reported for each fold. The model achieved consistent and robust performance in the segmentation task, with slightly higher Dice scores than those previously reported for the same dataset in the literature [11]. Conversely, landmark prediction proved to be more challenging and exhibited greater variability.

This aspect was also evident from the network's output behavior (see Fig. 2), for both tasks. Specifically, for this

Table 1. Training statistics, in terms of runtime, epoch at which the best model was saved, and its corresponding loss, are reported for each fold.

Fold	Runtime [s]	Epoch	Loss
Fold_1	16590	922	0.1041
Fold_2	16201	970	0.0780
Fold_3	16517	728	0.0648
Fold_4	16805	390	0.1420
Fold_5	16133	678	0.0834
Fold_6	17192	856	0.1002
Fold_7	16110	644	0.0789
Fold_8	15505	762	0.0508
Fold_9	16010	978	0.0919
Fold_10	16361	976	0.0754

Table 2. Performance metrics for the test set, in terms of Dice score and L2 distance, are reported for each fold.

	Dice [a.u.]		L2 [mm]	
	Img1	Img2	Img1	Img2
Average	0.9087		11.28	
Fold_1	0.9184	0.9624	8.76	3.04
Fold_2	0.9642	0.9013	22.37	6.62
Fold_3	0.6943	0.9545	19.83	9.45
Fold_4	0.9134	0.9014	8.07	18.04
Fold_5	0.9624	0.7691	13.18	12.86
Fold_6	0.9285	0.9538	5.84	4.27
Fold_7	0.9680	0.8973	10.21	23.63
Fold_8	0.9643	0.9026	14.24	17.10
Fold_9	0.9353	0.9028	7.04	6.29
Fold_10	0.8647	0.9156	11.41	3.32

image, the resolution is 0.365×0.365×0.625. The target coordinates in pixels are (388, 364, 72), while the network predicted the location at (372, 364, 72). Once we converted these coordinates in mm, we got an L2 distance of 5.84 mm, which reflected a suitable performance in the prediction of the landmark site.

A possible explanation of the non-negligible errors in the landmark prediction, given the overall mean distance of 11.28 mm, is that the network tends to predict a whole region of interest, rather than a specific point. Moreover, the manual process for annotating the apex could be

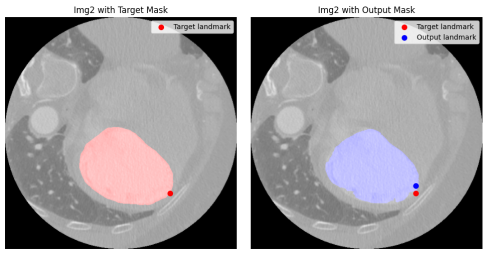

Figure 2. Performance of the network on a test image. On the left, a slice of the CT with the ground truth left ventricle mask in shaded red and the ground truth location of the apex, always in red. On the right, the same slice with the predicted mask in shaded blue and both the ground truth location of the apex (in red) and the predicted one (in blue).

biased by the single-expert annotation, and might be imprecise for some images, especially when the apex area is smooth and large. This finding, however, doesn't limit the application of this approach as a pre-processing step for image alignment, given the need for multiple landmarks. Finally, a notable limitation of this study is the dataset size, which restricts the generalizability of the results and led to suboptimal training of the network.

4. Conclusion

This study proved the feasibility of using a multi-task network for the segmentation of the left ventricle and the prediction of the coordinates of its apex. The proposed model achieved a high segmentation accuracy and a suitable landmark localization performance.

Results suggest that the proposed technique could serve as an effective preprocessing step for aligning the volumetric image of a cardiac chamber with a functional image where the same structures and points can be identified. This is particularly relevant for generating multimodal images in cardiac electrophysiological or structural studies, to automate the image fusion process.

Future developments will focus on a larger dataset, with a higher number of landmarks to perform accurate alignment. Landmark annotation by multiple experts could improve training and enhance network performance. Refinements in the data pre-processing stage will be considered, as well as potential variations in the network architecture to predict more precisely the landmark locations.

Acknowledgments

This work used the HPC DataCenter at IUSS, co-funded by Regione Lombardia through the funding programme established by Regional Decree No. 3776 of November 3, 2020.

References

[1] M. A. Daubert, T. Tailor, O. James, L. J. Shaw, P. S. Douglas, and L. Koweek, "Multimodality cardiac imaging in the 21st century: evolution, advances and future opportunities for innovation," *British Journal of Radiology*, vol. 94, no. 1117, p. 20200780, Jan. 2021, doi: 10.1259/bjr.20200780.

[2] A. P. Pazhenkottil *et al.*, "Prognostic value of cardiac hybrid imaging integrating single-photon emission computed tomography with coronary computed tomography angiography," *Eur Heart J*, vol. 32, no. 12, pp. 1465–1471, 2011, [Online]. Available: https://www.zora.uzh.ch/id/eprint/47543/

[3] D. Andreu *et al.*, "Integration of 3D electroanatomic maps and magnetic resonance scar characterization into the navigation system to guide ventricular tachycardia ablation," *Circ Arrhythm Electrophysiol*, vol. 4, no. 5, pp. 674–683, Oct. 2011, doi: 10.1161/CIRCEP.111.961946.

[4] H. B. Winther *et al.*, "v-net: deep learning for generalized biventricular mass and function parameters using multicenter cardiac MRI data," *JACC Cardiovasc Imaging*, vol. 11, no. 7, pp. 1036–1038, 2018, doi: https://doi.org/10.1016/j.jcmg.2017.11.013.

[5] S. Gao, H. Zhou, Y. Gao, and X. Zhuang, "BayeSeg: Bayesian modeling for medical image segmentation with interpretable generalizability," *Med Image Anal*, vol. 89, p. 102889, 2023, doi: https://doi.org/10.1016/j.media.2023.102889.

[6] X. Zhuang, "Multivariate mixture model for myocardial segmentation combining multi-source images," *IEEE Trans Pattern Anal Mach Intell*, vol. 41, no. 12, pp. 2933–2946, 2019, doi: 10.1109/TPAMI.2018.2869576.

[7] X. Luo and X. Zhuang, "X-Metric: an N-dimensional information-theoretic framework for groupwise registration and deep combined computing," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 7, pp. 9206–9224, 2023, doi: 10.1109/TPAMI.2022.3225418.

[8] F. Wu and X. Zhuang, "Minimizing estimated risks on unlabeled data: a new formulation for semi-supervised medical image segmentation," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 5, pp. 6021–6036, 2023, doi: 10.1109/TPAMI.2022.3215186.

[9] P. A. Yushkevich *et al.*, "User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006, doi: https://doi.org/10.1016/j.neuroimage.2006.01.015.

[10] Z. Tan, J. Feng, W. Lu, Y. Yin, G. Yang, and J. Zhou, "Multi-task global optimization-based method for vascular landmark detection," *Computerized Medical Imaging and Graphics*, vol. 114, p. 102364, 2024, doi: https://doi.org/10.1016/j.compmedimag.2024.102364.

[11] C. Wang, T. MacGillivray, G. Macnaught, G. Yang, and D. Newby, "A two-stage 3D Unet framework for multi-class segmentation on full resolution image," *Computer Vision and Pattern Recognition*, 2018, doi: https://doi.org/10.48550/arXiv.1804.04341.

Address for correspondence:

Danilo Pani
MeDSP Lab, Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy
danilo.pani@unica.it