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 performs the segmentation of the left ventricle and the localization of its apex, and 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 11.28mm for the segmentation and the landmark detection, respectively. These results suggest that, with some improvements, the proposed technique could be used as a preprocessing step when aligning the volumetric image of a cardiac chamber to another structure.

1. Introduction

Cardiac imaging plays a crucial role in the diagnosis and treatment of cardiovascular diseases. However, different imaging techniques may provide different insights on cardiac anatomy and its functionalities. Indeed, structural imaging such as computed tomography (CT) or cardiac magnetic resonance (CMR) provides detailed anatomical information, but they can fail in fully capturing functional information. On the other hand, complementary functional techniques like the electroanatomical (EA) mapping may support in enhancing the cardiac electrophysiological information, even though with less accurate anatomical background. Hence, the integration of all this information into a single, multimodal image is often desired [1]. Indeed, 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, demonstrating 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. Consequently, the 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.

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.

Given the purpose of this study, only the labeled images were considered, for which the dataset provides manual segmentation of seven whole heart substructures, including the left ventricle blood cavity, which we focused on for the segmentation. To pursue the research goal, in addition to the provided segmentation, we added the annotation of a landmark indicating the apex of the left ventricle to each image of the selected dataset. This landmark was marked by an expert on CT scans via the ITK-SNAP application [9], in the form of 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, which was added with ITK-SNAP onto the mask. Then, a heatmap was obtained by applying the Euclidean distance transform, with zeros representing the landmark location, followed by a logarithmic transform to highlight the landmark more. 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. Specifically, sub-volumes of size 128×128×128 were extracted while ensuring a balance between positive and negative samples, where the positive ones were regions containing the target label (i.e., the left ventricle), while the negative samples represented background or non-target regions. Their ratio was set to 1:1.

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. Moving to 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 part of the network, indicated by the dashed lines in Fig. 1, preserve spatial information, which is crucial 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, resulting in sixteen images for training, two for validation, and two for the test set. We ensured that no repetitions were allowed, so that in each fold the testing subjects couldn't also be present in the training and/or validation set.

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 segmentation task and the landmark position prediction task, 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.

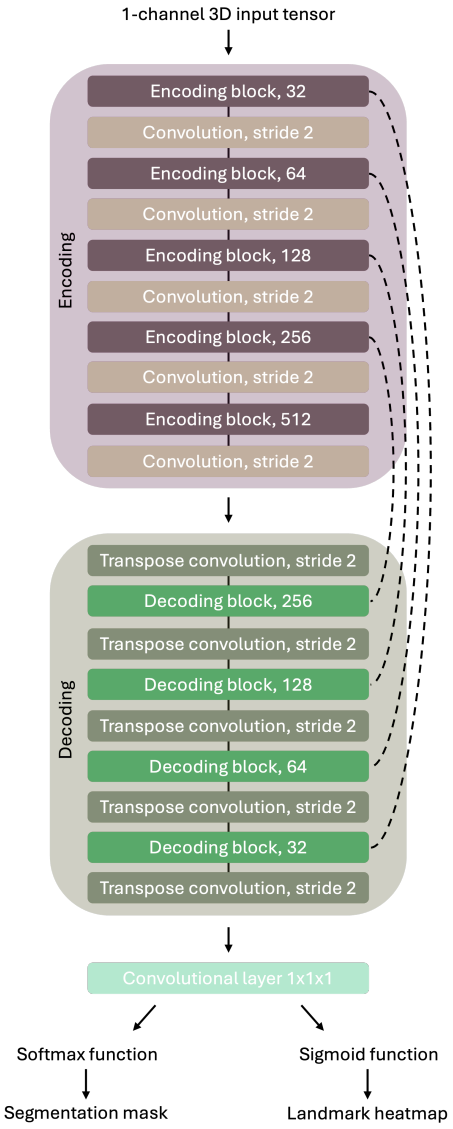

Figure 1. Model architecture, based on a 3D UNet. Input is represented at the top, followed by the encoding block (light purple), the decoding block (light green), the last convolutional layer, and the two outputs of the network. Skip connections between the encoding and decoding blocks are represented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

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 runtime for the training, the epoch at which the best model was saved, and the corresponding loss are reported for each fold. Overall, the runtime was 16342±466 s, and the lowest loss reached 0.0869±0.025, with the single values being consistent across folds.

In Table 2 the performance of the network on the images of the test sets are reported for each fold. These results indicate that the model achieved consistent and robust performance in the segmentation task, while landmark prediction proved to be more challenging and exhibited greater variability.

This aspect was also evident from the network's output behavior, an example of which is illustrated in Fig. 2, both for the segmentation and the apex detection tasks. Specifically, for this image, the resolution is 0.365×0.365×0.625. The target coordinates in pixels are (388, 364, 72), while the output of the network predicted the location at (372, 364, 72). Once we converted these coordinates in mm by multiplying them by their corresponding resolution, we got an L2 distance of 5.84 mm, which reflected a suitable performance in the prediction of the landmark site.

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		11.2795	
Fold_1	0.9184	0.9624	8.7636	3.0449
Fold_2	0.9642	0.9013	22.3705	6.6150
Fold_3	0.6943	0.9545	19.8319	9.4532
Fold_4	0.9134	0.9014	8.0671	18.0404
Fold_5	0.9624	0.7691	13.1799	12.8581
Fold_6	0.9285	0.9538	5.8437	4.2656
Fold_7	0.9680	0.8973	10.2139	23.6336
Fold_8	0.9643	0.9026	14.2407	17.1018
Fold_9	0.9353	0.9028	7.0416	6.2903
Fold_10	0.8647	0.9156	11.4108	3.3227

Although the landmark prediction results exhibit non-negligible errors, with an overall mean distance of 11.28 mm, some considerations can be drawn. A possible explanation is that the network might be more inclined to predict a whole region of interest, rather than a specific point. Moreover, the manual process for annotating the apex could be biased by the presence of a single expert 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 methodology 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 deep neural 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.

The 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 of the same cardiac structure. This capability is particularly relevant for generating multimodal images in cardiac electrophysiological or structural studies, to automate the image fusion process.

Future developments of this work will focus on a larger dataset, also considering a higher number of landmarks needed to perform an accurate alignment. Refinements in the data pre-processing stage will also be considered, as well as potential variations in the network architecture to predict the locations of the landmarks more precisely. Finally, annotating the landmarks by multiple experts could improve training and, consequently, enhance network performance.

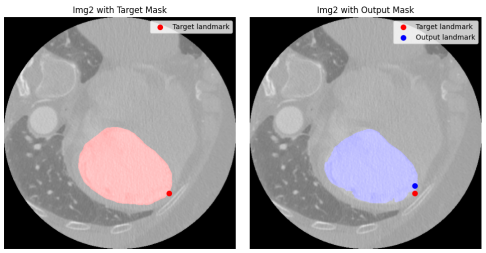

Figure 2. Performance of the network on a test image. On the left, a slice of the CT with the ground truth left ventricle mask in shaded red and the ground truth location of the apex, always in red. On the right, the same slice with the predicted mask in shaded blue and both the ground truth location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

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.

Address for correspondence:

Nicla Mandas
Istituto Universitario di Studi Superiori, IUSS, Pavia, Italy
MeDSP Lab, Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy
nicla.mandas@iusspavia.it