Evaluation of Anatomical Shape Priors in Deep Learning-Based Cardiac Multi-Compartment Segmentation

Michael Hudler, Franz Thaler, Martin Urschler
Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz
{michael.hudler, martin.urschler}@medunigraz.at

Abstract

Whole-heart multi-compartment CT segmentation is clinically important, but standard CNNs do not explicitly enforce anatomical plausibility. Based on statistics derived from the training data, we evaluate whether lightweight explicit shape priors, implemented as shape-aware losses and spatial label distribution heatmap-guided U-Net variants, improve 3D cardiac segmentation on MM-WHS CT and WHS++. Across all experiments, a standard 3D U-Net surprisingly remained a very strong baseline, with handcrafted priors yielding at best marginal and inconsistent changes and often degrading performance. These results suggest that the baseline already captures substantial implicit anatomical regularities and that future gains will likely require more expressive learned priors rather than simple handcrafted anatomical shape constraints.

1 Introduction

Multi-compartment whole-heart segmentation from cardiac CT [1] is a core task in medical image analysis because it supports quantitative assessment of clinical parameters like ejection fraction, builds the foundation for treatment planning or simulation [2], and enables image-guided interventions. Deep learning [3], especially 3D U-Net variants [4, 5], has become the dominant approach for this problem due to its strong multiscale feature extraction capabilities and its support for accurate localization [6]. Their success is explained by the clever combination of encoder-decoder feature extraction, multiscale context modeling, and skip connections that preserve spatial detail. However, these models are primarily appearance-driven and do not explicitly encode anatomical shape knowledge as was prominently done in the pre-deep learning era via statistical shape models [7, 8]. This gap motivates the study of shape priors in deep learning-based segmentation [9].

This work evaluates whether explicit shape priors improve whole-heart multi-compartment CT segmentation beyond a strong 3D U-Net baseline. Rather than tediously building a full statistical shape-modeling pipeline, our study tests lightweight priors that can be incorporated directly into the training objective or 3D network design. The central question is whether such priors provide measurable benefit on modern deep learning baselines for seven-class cardiac CT segmentation, involving ventricles, both atria, myocardium, and the great vessels. The main finding is negative but clear: in the studied setting, explicitly designed shape priors did not consistently improve performance over a well-trained 3D U-Net baseline.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

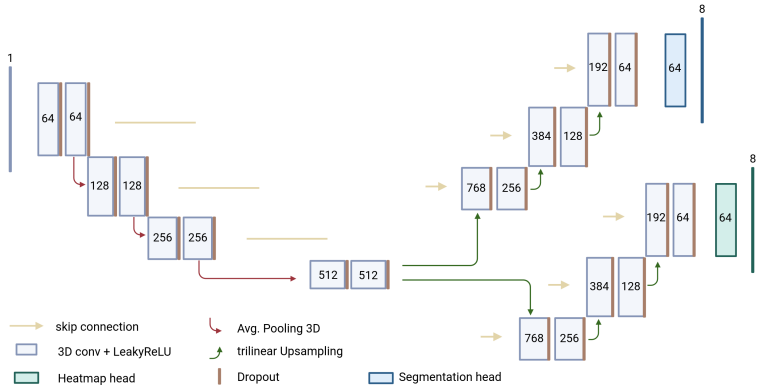

Figure 1: Exemplary architecture of one of our proposed networks incorporating label distribution heatmaps: The 2-Decoder network, with separate decoders for label segmentation and heatmap predictions, thus forcing the encoder to extract features supporting both predictions.

2 Methods

Dataset and preprocessing. Experiments are based primarily on the CT subset of the MM-WHS challenge dataset [1], comprising seven foreground classes: left ventricle, right ventricle, left atrium, right atrium, myocardium, ascending aorta, and pulmonary artery. The benchmark provides 20 annotated CT scans for training and 40 CT test scans evaluated with the official hidden-label evaluation script used for the Challenge. To complement this with accessible ground truth, evaluation was extended using the second half of the WHS++ training set (20 CT cases), which corresponds to the publicly released extension of the MM-WHS CT dataset. Images and labels were reoriented to a common anatomical convention, resampled isotropically, centered using label centroids, and embedded into a standardized field of view. A Procrustes-based alignment [10] of training labels was further used to derive population heatmaps representing average spatial label distributions in the registered space.

Baseline model. The reference model is a standard 3D U-Net [4, 5] with single-channel CT input and eight output classes including background. It uses a conventional encoder-decoder design with skip connections with 64 base channels and doubling the number of channels at each downsampling step. It uses LeakyReLU activations [11] and is trained with a combined Generalized Dice [12] and Cross-Entropy loss. This baseline serves as the main point of comparison throughout the study.

Shape-aware losses. Three families of explicit regularizers were evaluated in combination with the baseline loss. *Volume regularization* penalizes deviations from expected compartment volumes estimated from the training set via the label-specific volume means and standard deviations. Moment-based *shape regularization* compares soft first- and second-order spatial moments of predictions to reference shape moments (centroids, ellipsoids) from the training set via L2 distance. *Anatomical relation* loss constrains pairwise distances and angular relations between class centroids via reference angle statistics derived from the training data. All losses aim to inject coarse anatomical prior knowledge without changing the overall segmentation backbone.

Architectural priors. In addition to loss-based priors, we also investigated population-level multi-class probability heatmaps derived from aligned labels of the training dataset. These shape priors were integrated into several U-Net variants: a model with an auxiliary heatmap prediction head attached to the last decoder layer, a multilayer deep-supervision version of the latter architecture (*HM multilayer*), a two-decoder network with separate segmentation and heatmap branches (*2-Decoder*, see exemplary architecture depicted in Fig. 1), a dual-encoder network that processes image and heatmap inputs in parallel (*2-Encoder*), and a cascaded three-U-Net architecture for coarse prediction and refinement (*Cascaded*).

Experimental setup. Models were trained on cropped regions of interest at 64^3 and 128^3 input resolution using the same extensive geometric and intensity data augmentation for all architectures. Evaluation used Dice, Jaccard, Hausdorff distance (HD), and Average Symmetric Surface

Distance (ASSD), as overlap- and boundary-based metrics, respectively. Qualitative comparisons were additionally assessed on WHS++, since the MM-WHS test set did not provide ground truth segmentations.

3 Results

Table 1 summarizes the main findings. More results, additional descriptions of methods and implementation details can be found in [13]. On MM-WHS at 64^3, the baseline achieved 90.85% Dice, 83.63% Jaccard, 7.64 mm HD, and 1.03 mm ASSD. Volume and moment regularization were essentially tied with the baseline (90.85% and 90.84% Dice), whereas the anatomical relation loss reduced performance to 88.98% Dice. Thus, simple handcrafted losses did not harm but also did not improve the already strong baseline.

Table 1: Main quantitative results. MM-WHS values report Dice, Jaccard, HD, and ASSD; WHS++ reports Dice only, as summarized in the thesis. Best values per block are in bold.

Setting Method	Dice (%)	Jaccard (%)	HD (mm)	ASSD (mm)
MM-WHS CT, 64^3, shape-aware losses				
Baseline	**90.85**	**83.63**	**7.64**	**1.03**
Volume regularization	**90.85**	83.62	7.70	1.04
Moment regularization	90.84	83.60	7.67	**1.03**
Anatomical relation	88.98	80.65	8.23	1.27
MM-WHS CT, 64^3, selected architectural priors				
Baseline	**90.85**	**83.63**	7.64	**1.03**
HM multilayer	90.60	83.23	7.78	1.06
2-Decoder	90.73	83.43	7.58	1.06
Cascaded	90.32	82.74	**7.55**	1.08
MM-WHS CT, 128^3, selected architectural priors				
Baseline	**92.05**	**85.78**	7.35	**0.88**
HM multilayer	91.80	85.38	7.28	0.90
2-Encoder	92.02	85.70	7.40	0.89
Cascaded	92.04	85.70	**7.26**	0.89
WHS++ CT, 64^3, shape-aware losses				
Baseline	88.93	81.01	18.47	1.68
Volume regularization	89.09	81.26	**17.91**	1.63
Moment regularization	**89.16**	**81.47**	18.31	**1.61**
Anatomical relation	88.66	80.67	18.13	1.70
WHS++ CT, 64^3, selected architectural priors				
Baseline	**88.93**	**81.01**	**18.47**	**1.68**
HM multilayer	88.72	80.63	20.88	1.73
2-Encoder	87.75	79.43	18.52	1.77
Cascaded	86.13	77.05	21.27	2.09

At the architectural level, heatmap-guided models remained competitive but did not clearly surpass the reference U-Net. At 64^3, the 2-Decoder variant reached 90.73% Dice and the cascaded model produced the best HD (7.55 mm), suggesting slightly improved boundary refinement, but overall overlap remained below baseline. At 128^3, the baseline improved to 92.05% Dice, while the closest competitors, 2-Encoder and Cascaded, achieved 92.02% and 92.04%, respectively. Hence, higher resolution improved most models, but not our main finding.

Evaluation on WHS++ confirmed the same trend. The baseline achieved 88.93% Dice, while volume and moment regularization yielded only marginal changes, reaching 89.09% and 89.16% Dice, respectively. Moment regularization produced the best Dice, Jaccard, and ASSD, whereas volume regularization achieved the lowest HD (17.91 mm). However, these improvements were small and inconsistent. Architectural prior-based models did not outperform the baseline: HM multilayer and 2-Encoder showed slightly lower overlap, and the cascaded architecture degraded performance across all metrics. Overall, the main finding remained consistent across datasets and model families.

raw image | ground truth | baseline | shape loss | 2-Decoder

Figure 2: Representative qualitative comparison on WHS++ (subject 2014, coronal slice 76). The baseline U-Net, the best loss-based prior (Mean-Shape), and the best architecture-based prior (2-Decoder) all capture the global anatomy well. Differences are mainly confined to boundaries and smaller structures.

Qualitatively, all models reproduced the overall cardiac configuration and inter-structure arrangement well (Fig. 2). The shape-aware variants appeared visually very similar to the baseline, with differences concentrated at boundaries and in thinner structures rather than in gross anatomical localization. This visual pattern is consistent with the quantitative results: the baseline already learned strong global anatomical regularities, leaving limited room for coarse handcrafted priors to add useful information.

4 Discussion and Conclusions

The central result of this study is that, surprisingly, explicit handcrafted shape priors did not consistently outperform a strong 3D U-Net baseline for whole-heart CT segmentation, which is performing in-line with the winner of the MM-WHS Challenge (see [6, 1]) as well as the participants at the Challenge associated with WHS++ [14]. This is a relevant negative result. It suggests that on MM-WHS and WHS++, the baseline model already learns substantial implicit anatomical regularities directly from the image data, and that coarse constraints such as expected volumes, low-order moments, centroid relations, or average label distribution heatmaps add little information beyond that baseline.

Several factors likely explain this outcome. First, the tested priors describe anatomy only at a coarse level and cannot capture the complex nonlinear variability of cardiac shape. Second, the remaining errors are mainly boundary-related, whereas the priors regularize global structure more strongly than local boundary detail. Third, because baseline performance is already high, measurable improvements are inherently limited. The experiments also show that greater architectural complexity does not automatically improve segmentation: models such as 2-Encoder and Cascaded remained competitive, and Cascaded slightly improved HD, but none clearly surpassed the simpler baseline in overall Dice or boundary overlap.

In summary, this work provides a focused evaluation of explicit shape priors in deep learning-based whole-heart segmentation and shows that simple handcrafted priors are insufficient to reliably improve a strong 3D U-Net. Future work should therefore move toward more expressive learned anatomical priors, such as generative diffusion-based [15] or flow matching-based [16] models trained on segmentation masks, which may better represent the distribution of plausible cardiac anatomy.

Acknowledgments and Disclosure of Funding

This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/PAT1748423.

References

[1] Xiahai Zhuang, Lei Li, Christian Payer, Darko Štern, Martin Urschler, Mattias P Heinrich, Julien Oster, Chunliang Wang, Örjan Smedby, Cheng Bian, Xin Yang, Pheng-Ann Heng, Aliasghar Mortazi, Ulas Bagci, et al. Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge. *Medical Image Analysis*, 58:101537, December 2019.

[2] Elena Zappon, Luca Azzolin, Matthias A F Gsell, Franz Thaler, Anton J Prassl, Robert Arnold, Karli Gillette, Mohammadreza Kariman, Martin Manninger, Daniel Scherr, Aurel Neic, Martin

Urschler, Christoph M Augustin, Edward J Vigmond, and Gernot Plank. An efficient end-to-end computational framework for the generation of ECG calibrated volumetric models of human atrial electrophysiology. *Medical Image Analysis*, 107(Pt B):103822, October 2025.

[3] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, 2015.

[5] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Lecture notes in computer science, pages 424–432. Springer International Publishing, Cham, 2016.

[6] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Multi-label whole heart segmentation using CNNs and anatomical label configurations. In *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, pages 190–198. Springer International Publishing, 2018.

[7] Tim F Cootes, Chris J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

[8] Tobias Heimann and Hans-Peter Meinzer. Statistical shape models for 3D medical image segmentation: a review. *Medical Image Analysis*, 13(4):543–563, August 2009.

[9] Simon Bohlender, Ilkay Oksuz, and Anirban Mukhopadhyay. A survey on shape-constraint deep learning for medical image segmentation. *IEEE Reviews in Biomedical Engineering*, 16:225–240, January 2023.

[10] Peter H Schoenemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31:1–10, 1966.

[11] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning (ICML)*, volume 30, 2013.

[12] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support - DLMIA ML-CDS 2017*, volume 10553 of *Lecture Notes in Computer Science*, pages 240–248. Springer, September 2017.

[13] Michael Hudler. *Evaluation of Shape Models for Deep Learning Based Cardiac Image Segmentation*. Master's Thesis, Graz University of Technology, Graz, Austria, 2026.

[14] Franz Thaler, Darko Štern, Gernot Plank, and Martin Urschler. Augmentation-based domain generalization and joint training from multiple source domains for whole heart segmentation. In *Comprehensive Analysis and Computing of Real-World Medical Images. CARE 2024*, volume 15548 of *Lecture notes in computer science*, pages 168–179. Springer Nature Switzerland, Cham, 2025.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[16] Arnela Hadzic, Lea Bogensperger, Andrea Berghold, and Martin Urschler. Flow matching-based data synthesis for robust anatomical landmark localization. *IEEE Journal of Biomedical and Health Informatics*, 2025.