EXPLORING IMPROVEMENTS FOR MEDICAL IMAGING ON THE SEGTHOR DATASET

Pepijn de Reus[1], Didier Merk[1], Lisa van Ommen[1] and Raoul Ritter[1]

[1]Master Artificial Intelligence, Informatics Institute, Amsterdam, The Netherlands

1. INTRODUCTION

In 1895, Wilhelm Konrad Röntgen's discovery of X-rays laid the foundation for modern medical imaging [1]. Today, Computed Tomography (CT) scans are crucial for visualizing anatomical structures, particularly in the thoracic region. However, automatically segmenting different organs in these scans remains challenging due to varying organ shapes and similar tissue densities between structures.

In this paper, we address thoracic organ segmentation using the *__Seg__mentation of __Th__oracic __O__rgans at __R__isk* (SegTHOR) dataset [2], which provides CT scans and ground truth segmentations of four organs (esophagus, heart, trachea, and aorta) from 60 patients. Using the Efficient Neural Network (ENet) [3] as our baseline, we investigate four approaches to improve performance: preprocessing, data augmentation, hyperparameter tuning, and post-processing. We also compare ENet with recent architectures including Vision Mamba UNet (VM-UNet) [4] and Segment Anything Model 2 (SAM2) [5].

Our experiments show that preprocessing and model optimization significantly improve segmentation quality, achieving a 3D-Dice score of 0.8071 compared to the baseline's 0.6826. We evaluate performance using multiple metrics: 2D and 3D Dice Similarity Coefficient, 95th-percentile Hausdorff Distance, and Average Symmetric Surface Distance.

The remainder of this paper is organized as follows: Section 2 provides theoretical background, Section 3 describes our methodology, Section 4 presents experimental results, and Section 5 discusses our findings. We conclude with a summary and future directions.

2. THEORETICAL BACKGROUND

This section provides the theoretical foundation for our work through two main components. First, we introduce the model architectures employed in our experiments, detailing their key characteristics and design principles. Second, we discuss the the metrics used to evaluate segmentation quality, including their strengths and limitations.

2.1. Model Architectures

In this subsection, we present the three model architectures employed in our experiments: ENet, VM-UNet, and SAM2. Each model is designed to address the challenges of thoracic organ segmentation with varying degrees of complexity and computational efficiency.

2.1.1. ENet

The Efficient Neural Network (ENet) [3], released in 2016, is specifically designed for real-time semantic segmentation. Its architecture follows an encoder-decoder structure with particular emphasis on efficiency. The encoder consists of an initial stage and three bottleneck stages, that reduce the spatial dimensions and capture semantic features. The decoder part of the architecture is designed with two bottleneck stages to recover the original spatial resolution. The key features of the network are:

- Early *downsampling* through an initial block that combines max pooling and convolution
- An asymmetric encoder-decoder ratio (5:1) that prioritizes efficient feature extraction
- Modified residual blocks with PReLU activation functions
- Dilated convolutions in later stages for an enlarged receptive field
- Factorized filters ($1 \times n$ followed by $n \times 1$) reducing computational complexity

2.1.2. VM-UNet

Vision Mamba UNet (VM-UNet) [4] combines traditional convolutional approaches with state space models (SSMs). While the traditional Convolutional Neural Networks (CNNs) offer linear complexity, they struggle with long-range dependencies [4, 6]. Vision Transformers (ViTs) can capture these global dependencies better, however it comes at the cost of quadratic complexity and increased memory usage.

VM-UNet addresses these limitations by integrating UNet's [7] convolutional architecture with Mamba's [6] 2D-selective-scan (SS2D) method. This hybrid approach maintains linear complexity while effectively capturing global dependencies, resulting in competitive segmentation scores while maintaining computational efficiency [4].

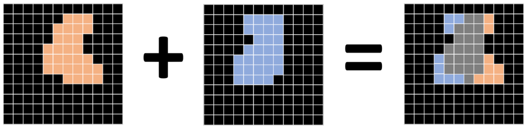

Fig. 1: Visualisation of a ground truth segmentation on the left and a segmentation prediction on the right. The amount of overlap is defined as the (in this case 2D) Dice Coefficient.

2.1.3. *SAM2*

Segment Anything Model 2 (SAM2) [5] is Meta's recent foundation model for segmentation tasks. It builds upon Hiera [8], a hierarchical vision transformer architecture that processes images at multiple resolutions to efficiently capture both local and global features. Trained on a diverse dataset of 51,000 videos containing 600,000 masks, SAM2 demonstrates strong zero-shot generalization capabilities across a wide range of segmentation tasks allowing for video and image input.

2.2. Evaluation Metrics

How do you measure performance in image segmentation? There are many different metrics to measure performance in image segmentation [9] and each of them captures different aspects of the performance. Measuring *accuracy* is different than measuring the *error*, for example. The orientation (2D or 3D) can also change the outcome of the segmentation metric.

Often, metrics are chosen inadequately, which impacts the model's ability to be applied in practice [10, 11]. To address this mismatch, we employ four distinct metrics to validate our segmentation model: 2D Dice Similarity Coefficient, 3D Dice Similarity Coefficient, 95th-percentile Hausdorff Distance, and Average Symmetric Surface Distance. Each metric provides unique insights into the model's performance, with the two Dice variants specifically helping us understand performance differences between slice-wise and volumetric evaluations.

2.2.1. *Dice Similarity Coefficient*

The Dice metric was originally drafted as a metric to compute ecologic association between species [12]. It gives an insight into how closely associated two surfaces (2D, Figure 1) or volumes (3D) are by calculating their amount of overlap.

2.2.2. *95th-percentile Hausdorff Distance*

Instead of looking at the amount of overlap between for example a ground truth and prediction segmentation, the Hausdorff Distance (or min-max) rather looks at the error of your prediction.

It is an inherently 3D metric, for each point on a prediction surface calculates the closest distance to the ground truth surface, and vice versa [13]. From all these "errors" it then takes the 95th-percentile distance, indicating that 95% of the two-volume boundaries are closer together than that distance.

This metric is susceptible to outliers, and ignores holes in surfaces and volumes [11]. Therefore this metric is often used for the detection of spatial outliers.

2.2.3. *Average Symmetric Surface Distance*

The Average Symmetric Surface Distance (ASSD) is very similar to the Hausdorff distance but instead of using the 95th-percentile it uses the average off all surface distances.

$$\text{ASSD}(A, B) = \frac{\sum_{a\in A} d(a, B) + \sum_{b\in B} d(b, A)}{|A| + |B|} \quad (1)$$

Equation 1 shows that the ASSD calculates the minimum distances between the surfaces for all points on the surfaces ($d(a, B)$ and $d(B, a)$), and then averages it by dividing by the total amount of points on the surfaces. This averaging operation makes the ASSD more robust to outliers and gives an idea of the average "error" of your model.

3. METHODOLOGY

This section details our experimental approach to improving medical image segmentation. All code and implementation details are publicly available on GitHub, enabling full reproducibility of our results.

Our experimental framework systematically evaluates various optimization techniques for the ENet baseline architecture through an ablation study. Each enhancement technique can be independently enabled or disabled, allowing us to isolate their individual impacts and identify optimal combinations for thoracic organ segmentation.

3.1. Data Preprocessing

All our experiments are run using the SegTHOR dataset [2]. As mentioned before, this dataset consists of the full-body CT scans and corresponding ground truth segmentations for 60 patients, both in *NIfTI* format. However, the quality of these CT scans is not always perfect or consistent across patients. Therefore, we apply multiple preprocessing steps to improve the quality of data we feed into the network.

3.1.1. *Heart Segmentation Correction*

Notably, the dataset contains a systematic discrepancy, causing all ground truth heart segmentations to be placed in the wrong position. The first step of the data preprocessing consists of fixing these segmentations. We can do this using patient number 27, the only patient with both a correct and corrupted ground truth. This allows us to calculate the corrupted

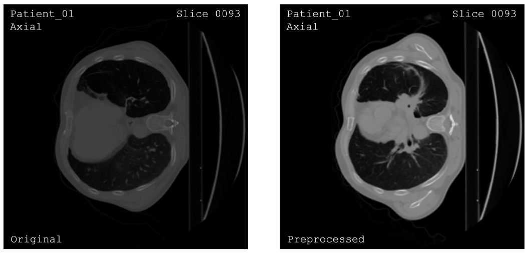

Fig. 2: Visualisation of a preprocessed CT-scan slice Here we applied voxel clipping, rescaling, and intensity normalization in order to improve the image quality.

and correct ground truth segmentation centroids. Using these centroids we can determine an affine transformation, to move all the heart segmentations for the other patients to their correct position. This is a vital step in order to learn meaningful segmentations during training.

3.1.2. Voxel Clipping

To reduce noise in CT scans, we use *voxel clipping*, which limits extreme voxel intensities by clipping them to a predefined threshold. In medical imaging, voxel intensities are measured in Hounsfield Units (HU), with air (-1000HU) and bone (1000HU) often representing the extremes. Therefore voxel intensities outside this range usually indicate noise or irrelevant details. By voxel clipping we focus the analysis on the most meaningful structures in the CT-scan, to enhance the model's ability to learn relevant features.

3.1.3. Rescaling

To standardize the spatial dimensions of all scans, we rescale the voxel size to $0.977\text{mm} \times 0.977\text{mm} \times 2.500\text{mm}$. This ensures that images have consistent voxel spacing, regardless of the original scan parameters. By making the voxel dimensions uniform, we reduce variability and improve the comparability of anatomical structures across different patients.

3.1.4. Intensity Normalization

The final preprocessing step that we perform is intensity normalization. This is done to account for differences in voxel intensity distributions across scans, which could arise from variations in scanners or imaging protocols. We use z-score normalization, calculated as $z = \frac{x-\mu}{\sigma}$, where x is the original voxel intensity, μ is the mean intensity, and σ is the standard deviation over all intensities. This process enhances contrast and brings all scans into a similar intensity range, improving the model's ability to generalize across different patients.

3.2. Models

We use ENet as our base model. Given the promising results in related work, we extend our evaluation to include VM-UNet and SAM2. All models are used with their default configurations, with minor modifications to VM-UNet and SAM2 to accommodate the grayscale images from the SegTHOR dataset. Our implementation is available on GitHub.

3.3. Data Augmentation

To improve generalization and increase the dataset size, we apply offline data augmentation by performing random affine transformations on each image slice, effectively doubling the dataset. These transformations are captured by the following affine matrix:

$$T = \begin{bmatrix} a_{11} & a_{12} & t_x \\ a_{21} & a_{22} & t_y \\ 0 & 0 & 1 \end{bmatrix}$$

We apply random rotations ($-10^\circ < \theta < 10^\circ$), uniform scaling ($0.9 < a_{11} = a_{22} < 1.1$), and translations ($-10 < t_x, t_y < 10$). These small adjustments simulate realistic variations in CT scans, allowing the model to train on more diverse data and reducing overfitting.

3.4. Hyperparameter Optimization

We performed hyperparameter optimization using Weights and Biases (wandb) sweeps with Bayesian optimization, efficiently exploring the hyperparameter space based on prior results. Our optimization focused on five key parameters: the learning rate (ranging from 1e-4 to 1e-2), optimizer (Adam vs. AdamW), learning rate scheduler (different modes with varying factors and patience values), weight decay (0.0001 to 0.01), and kernel size (16 to 64).

AdamW performed better than Adam, with its decoupled weight decay leading to improved generalization in our context. The optimal hyperparameters found through Bayesian optimization are shown below.

Hyperparameter	**Optimal Value**
`Learning Rate`	`1e-3`
`Optimizer`	`AdamW`
`LR Scheduler`	`mode='min', factor=0.5, patience=5`
`Weight Decay`	`0.0001`
`Kernel Size`	`32`

Table 1: Optimal hyperparameters values found

3.5. Post-processing

To further improve the segmentation predictions, we apply a three-step post-processing pipeline. First, we use binary closing to fill small holes within the predicted masks. Next, binary opening is applied to remove noise and small spurious regions. Finally, we retain only the largest connected component to focus on the primary structure of interest. This process ensures cleaner and more accurate segmentation results by eliminating irrelevant and noisy parts of the prediction segmentations. The ability to turn on or off all these settings individually allows us to train 13 different model variations of the ENet model, and compare them with each other and the VM-UNet and SAM2 models.

4. RESULTS

Our experiments evaluated model performance using 3D-Dice Similarity Coefficient (Dice), 95th-percentile Hausdorff Distance (HD95), and Average Symmetric Surface Distance (ASSD). Table 2 summarizes the results across configurations, while Figure 3 illustrates the performance distributions.

Preprocessing showed the most significant impact, improving the Dice score from 0.6826 (baseline) to 0.7576, while substantially reducing HD95 and ASSD metrics. The boxplots in Figure 3 demonstrate that this improvement is consistent across all organs, with notably reduced variance in the metrics after preprocessing.

Data augmentation further enhanced performance, particularly when combined with hyperparameter tuning, achieving our highest Dice score of 0.8071. The distribution plots show this configuration also provided the most stable performance across different organs. Interestingly, while post-processing produced visually appealing results, it often decreased metric scores, likely due to over-aggressive filtering of small structures.

The optimal configuration combined preprocessing, augmentation, and tuning using the AdamW optimizer, achieving the best scores across all three metrics. This is reflected in both the numerical results and the tighter distributions shown in the boxplots, particularly for HD95 and ASSD metrics. For visual comparison between our baseline and best model predictions against ground truth, see Figure 4 in the Appendix.

5. DISCUSSION

Our experimental results reveal varying segmentation performance across different organs, with the esophagus proving particularly challenging. This lower performance for esophageal segmentation can be attributed to two main factors: first, its anatomical characteristics as a small, elongated organ close to the aorta, and second, its inherent difficulty in medical imaging, as noted by related work stating that "The esophagus is one of the most difficult OARs to segment" [14].

We implemented data augmentation strategies to address this challenge with esophagus segmentation, applying up to three transformations per slice. We deliberately excluded mirroring transformations, as such configurations would be physiologically impossible in thoracic scans. While additional augmentations could have been generated through multiple transformation rounds, we determined that a single round sufficiently demonstrated the technique's value while balancing computational constraints.

Settings	**ENet (Adam)**			**ENet (AdamW)**		
	Dice	**HD95**	**ASSD**	**Dice**	**HD95**	**ASSD**
Baseline	0.6826	12.2335	3.5179	-	-	-
Preprocessed	0.7576	6.6451	2.2200	-	-	-
PreP + Tuning	0.7861	6.1526	1.9648	0.7752	5.8931	1.9484
PreP + Augmentation	0.8023	6.7640	1.7883	-	-	-
PreP + PostProcess	0.7528	14.0171	3.0799	-	-	-
PreP + Aug + Tune	0.7987	5.9352	1.9050	**0.8071**	**5.0892**	**1.6597**
PreP + Tune + PostP	0.7788	12.8022	2.8587	0.7699	11.7992	2.7076
PreP + Aug + PostP	0.8027	8.2353	2.1105	-	-	-
PreP + Aug + Tune + PostP	0.7844	13.5581	2.8232	0.8025	9.6998	2.2646

Table 2: Results for ENet (Adam) and ENet (AdamW) across different settings, with Dice, HD95, and ASSD as metrics. The best model for each metric is reported in **bold** and the second best for each metric is underlined.

Our experiments with state-of-the-art architectures revealed several significant challenges. VM-UNet, despite its theoretical advantages in capturing long-range dependencies through state space models, performed unexpectedly poorly on our segmentation task. This underperformance might be attributed to the model's sensitivity to hyperparameter settings and the specific characteristics of medical imaging data, which differ substantially from the natural images for which these architectures were primarily designed.

The fundamental design of SAM2 presented two major challenges for our task. First, its interactive segmentation approach expects prompting points to indicate target objects, which our dataset doesn't provide and proved difficult to generate automatically for specific organs. Second, SAM2's architecture is optimized for RGB images, requiring substantial modifications for grayscale medical imaging data. While we successfully implemented these modifications, these architectural challenges prevented us from completing our full evaluation suite within the project timeframe and computation budget.

5.1. Limitations

Reinke et al. (2024) [11] report that voxel-based metrics are not appropriate for detection problems. E.g. when detecting lesions after traumatic injury, voxel-based metrics miss the lesion. We deployed 3D-metrics as well but deem this limitation beyond the scope of our project. However, this limitation should be considered when discussing our framework's generalisability.

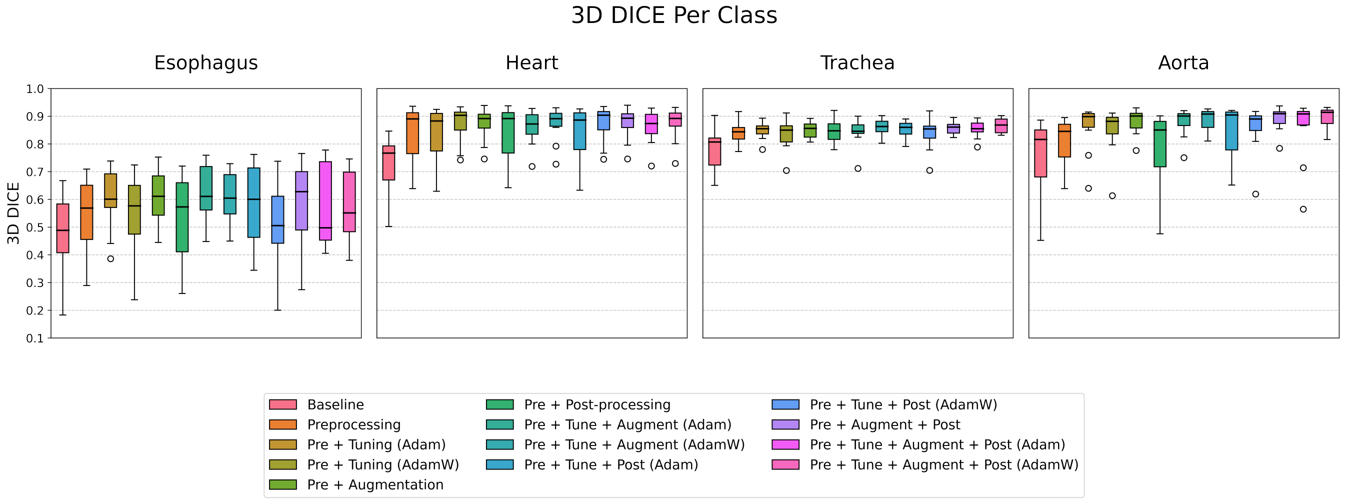

Fig. 3: Boxplot comparisons of model performance across different settings for the Dice, HD95, and ASSD metrics. The settings include variations in preprocessing, tuning, augmentation, and post-processing for both Adam and AdamW optimizers. These visualizations illustrate the distribution of performance metrics for each configuration, helping to identify the optimal combinations for thoracic organ segmentation.

5.2. Future work

A promising direction for future research lies in leveraging the temporal coherence inherent in 3D medical scans. While our current approach treats each slice independently, medical volumes can be viewed as a sequence of frames, similar to video data. This perspective enables several potential advances in medical imaging segmentation: treating consecutive CT slices as video frames to learn spatial continuity, propagating organ segmentations through volumes for improved consistency, utilizing previous slice information for automated prompting, and adapting SAM2's temporal tracking to follow organ boundaries continuously. These approaches could significantly reduce the current limitations of slice-by-slice processing while improving overall segmentation consistency.

6. CONCLUSION

In this work, we investigated approaches to improve medical image segmentation using the Segmentation of THoracic Organs at Risk (SegTHOR) dataset, which comprises CT scans of thoracic organs from 60 patients. Using the Efficient Neural Network (ENet) as our baseline architecture, we systematically evaluated multiple optimization strategies and compared their performance against state-of-the-art models.

Our experimental results demonstrate that preprocessing techniques showed the most significant impact on segmentation performance, while data augmentation and hyperparameter tuning provided additional performance gains. The combination of all optimization strategies achieved our best performance, with a 3D-Dice score of 0.8071 (0.9031 in 2D-Dice). Notably, despite their theoretical advantages, newer architectures (SAM2 and VM-UNet) did not outperform our optimized ENet baseline.

These results suggest that careful optimization of established architectures can match or exceed the performance of more complex, state-of-the-art models for specialized medical imaging tasks. Furthermore, our findings highlight the importance of preprocessing in medical image segmentation, indicating that data quality and standardization may be as crucial as model architecture choice, while future work could explore leveraging temporal coherence in 3D scans to further improve performance.

6.1. Report Requirements

The work was divided among the authors as follows: Lisa van Ommen focused on data preprocessing and augmentation strategies, Pepijn de Reus implemented VM-UNet and led paper writing, Didier Merk developed the core experimental framework and metrics, while Raoul Ritter handled hyperparameter optimization and SAM2 implementation.

7. REFERENCES

[1] P. Suetens, *Fundamentals of medical imaging*. Cambridge university press, 2017.

[2] Z. Lambert, C. Petitjean, B. Dubray, and S. Kuan, “Segthor: Segmentation of thoracic organs at risk in ct images,” in *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2020, pp. 1–6.

[3] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, “Enet: A deep neural network architecture for real-time semantic segmentation,” *arXiv preprint arXiv:1606.02147*, 2016.

[4] J. Ruan and S. Xiang, “Vm-unet: Vision mamba unet for medical image segmentation,” *arXiv preprint arXiv:2402.02491*, 2024.

[5] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, “Sam 2: Segment anything in images and videos,” *arXiv preprint arXiv:2408.00714*, 2024.

[6] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” *arXiv preprint arXiv:2312.00752*, 2023.

[7] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[8] C. Ryali, Y.-T. Hu, D. Bolya, C. Wei, H. Fan, P.-Y. Huang, V. Aggarwal, A. Chowdhury, O. Poursaeed, J. Hoffman, J. Malik, Y. Li, and C. Feichtenhofer, “Hiera: A hierarchical vision transformer without the bells-and-whistles,” 2023. [Online]. Available: https://arxiv.org/abs/2306.00989

[9] L. Maier-Hein, A. Reinke, P. Godau, M. D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek, M. Reyes, M. A. Riegler, M. Wiesenfarth, A. E. Kavur, C. H. Sudre, M. Baumgartner, M. Eisenmann, D. Heckmann-Nötzel, T. Rädsch, L. Acion, M. Antonelli, T. Arbel, S. Bakas, A. Benis, M. B. Blaschko, M. J. Cardoso, V. Cheplygina, B. A. Cimini, G. S. Collins, K. Farahani, L. Ferrer, A. Galdran, B. van Ginneken, R. Haase, D. A. Hashimoto, M. M. Hoffman, M. Huisman, P. Jannin, C. E. Kahn, D. Kainmueller, B. Kainz, A. Karargyris, A. Karthikesalingam, F. Kofler, A. Kopp-Schneider, A. Kreshuk, T. Kurc, B. A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A. L. Martel, P. Mattson, E. Meijering, B. Menze, K. G. M. Moons, H. Müller, B. Nichyporuk, F. Nickel, J. Petersen, N. Rajpoot, N. Rieke, J. Saez-Rodriguez, C. I. Sánchez, S. Shetty, M. van Smeden, R. M. Summers, A. A. Taha, A. Tiulpin, S. A. Tsaftaris, B. Van Calster, G. Varoquaux, and P. F. Jäger, “Metrics reloaded: recommendations for image analysis validation,” *Nature Methods*, vol. 21, no. 2, p. 195–212, Feb. 2024. [Online]. Available: http://dx.doi.org/10.1038/s41592-023-02151-z

[10] F. Kofler, I. Ezhov, F. Isensee, F. Balsiger, C. Berger, M. Koerner, B. Demiray, J. Rackerseder, J. Paetzold, H. Li *et al.*, “Are we using appropriate segmentation metrics? identifying correlates of human expert perception for cnn training beyond rolling the dice coefficient,” *arXiv preprint arXiv:2103.06205*, 2021.

[11] A. Reinke, M. D. Tizabi, M. Baumgartner, M. Eisenmann, D. Heckmann-Nötzel, A. E. Kavur, T. Rädsch, C. H. Sudre, L. Acion, M. Antonelli *et al.*, “Understanding metric-related pitfalls in image analysis validation,” *Nature methods*, vol. 21, no. 2, pp. 182–194, 2024.

[12] L. R. Dice, “Measures of the amount of ecologic association between species,” *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[13] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, “Comparing images using the hausdorff distance,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[14] J. O. B. Diniz, J. L. Ferreira, P. H. B. Diniz, A. C. Silva, and A. C. de Paiva, “Esophagus segmentation from planning ct images using an atlas-based deep learning approach,” *Computer Methods and Programs in Biomedicine*, vol. 197, p. 105685, 2020.

Appendix

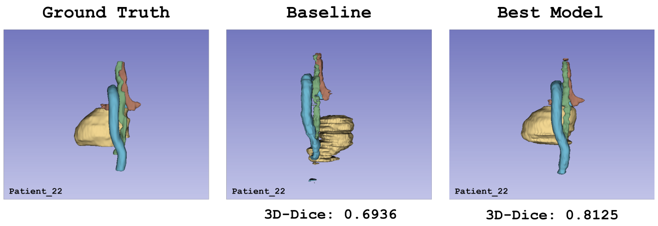

Fig. 4: Visualization of segmented thoracic organs for Patient 22, comparing the best model (right) with the baseline model (middle) and the ground truth (left). The segmented organs are the **esophagus**, **trachea**, **aorta**, and **heart**.