STUDY ON THE SEGTHOR CHALLENGE

Alin Prundeanu[1], Anesa Ibrahimi[1], Leon Hulshof[1],Max de Redelijkheid[1], Thomas Komen[1]

[1]Universiteit van Amsterdam, Amsterdam, The Netherlands

1. INTRODUCTION

For lung and esophageal cancer, radiation therapy is one of the standard treatments [1]. When applying radiotherapy to a tumor in this region it is important that certain organs are not irradiated. This is why during the planning stage, CT scans are manually annotated by specialists who are relying on their expertise and certain guidelines, which is still a time-consuming and tedious task. Data science algorithms, and by extension neural network algorithms, can play a big role in automatically detecting these areas to make radiotherapy as precise and effective as possible. To this end, the SegTHOR dataset was created which provides manually annotated CT scans of 60 patients, segmenting the most important organs at risk (OARs) in the thoracic region of the body. This dataset can be used to train a neural network for automatically segmenting these organs on new patients.

The goal of this paper is to compare different variations of neural methods trained for automatically segmenting OARs in the thoracic region of the human body. Due to the relatively small size of the data, we look at CNN-like deep neural networks which are less expensive to train compared to newer transformer-like methods. In this family of networks, previous literature suggests that ENet [2] performs (close to) SOTA in terms of performance and compute time [3, 4, 5] on several medical image segmentation tasks, which is why we use ENet as a baseline for our experiments.

Starting with a basic implementation of ENet, we try to improve its performance on SegTHOR by testing different combinations of loss functions and hyperparameters. Secondly, we implement several popular (CNN-like) methods and compare their outcome to ENet. This paper also highlights some shortcomings of the original SegTHOR dataset and ways to mitigate these limitations.

2. THE SEGTHOR DATASET

The SegTHOR (Segmentation of THoracic Organs at Risk) dataset contains 3D images with segmentations of organs at risk (OARs), specifically in the thorax region [1]. The full dataset contains 3D CT scans of 60 cancer patients acquired at the Henri Becquerel Center (CHB), with annotations for the heart, trachea, aorta and esophagus. Because this data is manually annotated, there are inconsistencies in the alignment and coverage of some annotations which will be discussed in chapter 3.1.

In the original paper, the creators of the SegTHOR dataset note several challenges with the manual annotations. Firstly, due to a lack of contrast, some parts of annotations rely on the anatomical knowledge of the expert instead of the actual CT data. Secondly, due to the nature of these organs, their segmentations are intricately interlocked. Finally, the shape and size of each organ and thus their segmentations differ vastly between organs, of which the esophagus is almost invisible in some patients.

An analysis of the data has shown a large class imbalance for the different organs segmented in the challenge. Appendix A contains a table with the total amount of images for each organ in the train and validation set, and their relative size. One issue shown here is the underrepresentation of the heart compared to other organs while having a relatively larger area. Some loss functions struggle with punishing mistakes with large segments, reducing the effectiveness of the model in learning about the shape of the heart [6].

3. METHODOLOGY

3.1. Data pre-processing and augmentation

By observing the data, it becomes apparent that the 3D segmentations have an inconsistency issue with the heart segments. The heart is disconnected from the aorta and is not placed in the correct location with regards to the CT scans.
One of the patients present in the data contained two segmentations, one of the annotations containing the original location of the segments and the other containing the corrected position for the heart. Using the latter as a ground truth, we applied automatic image registration from the elastix toolbox [7, 8] in 3D Slicer [9] to obtain the affine transformation required to position the heart correctly. As the data contained only one comparison, the matrix obtained from this pair is used to reposition the hearts for all patients present in the data. The CT scans and the edited 3D segmentations are then sliced to obtain 2D scans on the axial plane to accommodate the architecture of the networks used for segmentation. As these slices could contain fully empty segmentations when none of the annotated organs are present, we experimented with removing all scans with empty segmentations from training.

Following previous research on SegTHOR, we applied normalization to the CT scans to improve image contrast [1, 10] and a Gaussian blur to improve model robustness [10, 11].

3.2. Loss functions

Various loss functions can be used to effectively train vision models, particularly in medical imaging where accurate segmentation of anatomical structures is crucial. Our baseline ENet uses a basic cross-entropy loss due to its efficiency in classification tasks, however cross-entropy can struggle with imbalanced datasets where certain classes are underrepresented. Given the significant class imbalance in our dataset, several alternative loss functions are tested to improve segmentation accuracy.

Dice similarity coefficient loss or commonly referred to as Dice loss, measures the overlap between ground truth and predicted regions by directly optimizing the Dice coefficient [12]. It is particularly effective for imbalanced tasks, emphasizing the correct prediction of smaller, underrepresented regions.

Combined loss; To address the limitations of single loss functions, a combined loss was assessed, combining the cross-entropy and Dice loss with the hyperparameters α and β: $\boldsymbol{CL = \alpha \cdot CE + \beta \cdot DL}$. This approach utilizes the stability of cross-entropy and the imbalance handling of Dice loss, enhancing overall segmentation [13].

Focal loss builds upon cross-entropy by putting more emphasis on hard-to-classify examples, effectively handling class imbalance by focusing training on underrepresented regions [14]. This improves segmentation performance on smaller classes.

Tversky loss builds forth upon Dice loss to control the balance between false positives and false negatives, making it suitable for precision-recall trade-offs [15]. By adjusting weights, Tversky loss prioritizes smaller regions like the esophagus and trachea, achieving a more balanced segmentation across classes.

3.3. Base model - ENet

The convenience of having the codebase already provided helped us pick ENet as the baseline model for our experiments. Subsequently, to thoroughly assess the capabilities and limitations of ENet in our domain, various training and model configurations have been explored. In the end, two versions of ENet have been further used in our reported results: baseline and its final optimized variant.

Baseline-ENet is based on the original ENet implementation [2], while also keeping the training setup simple. Therefore, the model architecture and its intrinsic parameters are identical, while training is performed for 200 epochs with Adam optimizer and Cross-Entropy loss.

Best-ENet is fine-tuned on our given domain, after a grid search for finding suitable hyper-parameters. Considering the limited dataset and the signs of overfitting observed on the Baseline-ENet during training, additional regularization methods have been applied. As such, the model's dropout value for each layer after bottleneck2.0 is increased to 0.2 (from 0.1 in the baseline), and AdamW optimizer is used with a weight decay value of 0.1, while training is executed for 100 epochs with a OneCycle learning scheduler and an early stopping criterion of 10 consecutive epochs without an improvement on the validation Dice score.

3.4. Different architectures

To contextualize the performance of ENet, additional architectures have been explored and evaluated in our thoracic segmentation challenge. Specifically, three models have been picked, based on the literature for the commonly used deep learning algorithms in medical image segmentation [16]. These models employed the same hyper-parameter configuration as Best-ENet, after an initial random search showed no improvement from altering these values. Furthermore, we probed different encoders within the ResNet family and selected ResNet18 as the suitable backbone due to its favorable performance and small size compared to other variants. The encoder was initialized with ImageNet pre-trained weights, keeping the last layer unfrozen, which leverages transfer learning for the general image processing capabilities, while also fine-tuning for our specific domain knowledge.

UNet has been investigated due to its robust presence in classic literature, especially in medical image segmentation tasks. As described in the original paper [17], it is a fully convolutional neural network characterized by its symmetric U-shaped structure with encoder-decoder paths and skip connections that enable the capture of both contextual and spatial information.

UNetPlusPlus builds on top of UNet by incorporating nested and dense skip connections [18], reducing the semantic gap between the encoder and decoder feature maps, and thus enhancing the network's ability to capture fine-grained details and improving segmentation performance, especially in our scenario with limited annotated data.

DeepLabV3Plus differs by utilizing atrous spatial pyramid pooling and atrous (dilated) separable convolutions in both the encoder and decoder [19]. This enables it to control the receptive field size and effectively capture multiscale contextual information without significantly increasing computational cost.

3.5. Metrics

In order to measure and compare our different approaches for the image segmentation task, we employed several metrics. As mentioned in [20], the ideal metrics are required to fulfill several criteria, starting with identifying the differences between segmentation and the corresponding ground truth. Additionally, these metrics should be able to indicate different

types of segmentation errors (size, location, and shape), and serve as a ranking tool for evaluating different segmentation algorithms. As such, we decided to use **six** main evaluation metrics. Two of these metrics focus on measuring *similarity* between the segmented output and the ground truth: the Dice Similarity Coefficient (DSC) and the Jaccard Similarity Coefficient (JSC). The DSC metric is one of the most widely used metrics in medical image segmentation and it generally provides more weight to the correct predictions, while also punishing the FP (*false positives*) to handle the high-class imbalance in datasets [21]. The JSC metric [20], while similar to DSC, provides a slightly different perspective by penalizing FP and FN (*false negatives*) even more drastically than Dice.

Additionally, we decided to include two opposing metrics that instead serve as *difference* metrics: Average Hausdorff Distance (AHD), and Average Symmetric Surface Distance (ASSD). The AHD metric [21] provides a stable measurement of surface-to-surface similarity between the predicted and ground truth contours while measuring the maximum distance from any point on the boundary of the segmented region to the nearest point on the ground truth boundary and vice versa, ensuring that every boundary point is within a specified distance. The ASSD metric [20] calculates the average distance between the surfaces of the predicted segmentation and the ground truth in both directions. Lastly, we implemented Precision and Recall [22], which are essential in medical image segmentation tasks to grasp a more detailed view of model's performance in terms of accurately identifying the anatomical regions of interest without including other non-important regions (precision) and without missing any point from the area of interest (recall).

Moreover, the units used for the distance evaluation metrics (AHD & ASSD) are in **pixel space** as opposed to millimeters or other real-world units, which could influence the generalization of the resulting measurements. Furthermore, all these metrics are calculated as micro-averages over 2D slices from the SegTHOR dataset. However, given that 3D reconstructions were also created using 3D viewer [23], this enabled the evaluation of the segmentations in 3D space as well. Therefore, a 3D DSC was also implemented which instead performs *voxel-wise* comparison on entire 3D volumes of a patient, considering the entire structure when calculating the Dice score. When utilizing the AHD metric, certain adjustments were necessary to handle discrepancies between the predicted and ground truth boundaries. Firstly, zero-length boundaries were introduced when either the predicted or ground truth mask was empty. Additionally, to address large Hausdorff distances, a threshold of 1000 (a hyperparameter) was defined to limit excessively large values.

4. RESULTS

In this section, we first present our investigations on how different loss functions and data pre-processing&augmentation techniques affect the outcome of Baseline-ENet. Then, incorporating the best-performing techniques into the final four models, we report their performance on the entire suite of metrics presented above. All of our experiments have been realized in Pytorch and can be found in our public repository [1]. The three non-ENet final models are implemented using the Segmentation Models library [2].

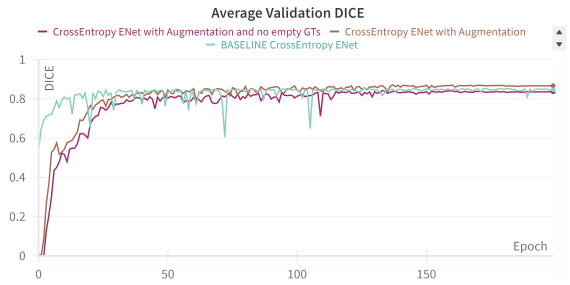


Fig. 1: Baseline-ENet using the data augmentation improves the validation DSC from 0.846 to 0.867, while removal of the empty ground truths decreases this to 0.843.

4.1. Data pre-processing and augmentation

Running Baseline-ENet without and with heart transformation shows an improvement in the Dice score from 0.846 to 0.869. Figure 1 shows the baseline ENet running with Augmentation and the removed ground truths. It can be observed that the augmentations improve baseline ENet from 0.846 to 0.867. Removing the empty ground truths on top of the augmentations decreases the performance to 0.843, but the evolution of the score remains the same.

4.2. Comparison of ENet loss functions

To find the best loss function for training our models, we compared the training performance of Baseline-ENet for all loss functions using the Dice Score Coefficient (DSC). Figure 2 shows the evolution of the DSC on the validation set during the training of the loss functions with their best-performing hyperparameters selected. Although the results between the cross-entropy, focal, and combined loss are close, the combined loss with $\alpha = .7$ and $\beta = .3$ consistently shows the highest performance during training.

We also observed that the Dice loss and the Tversky loss flatline during training. The Tversky loss flatlines at 12

[1] https://github.com/prundeanualin/ai4mi_project

[2] https://segmentation-models-pytorch.readthedocs.io/en/latest

Table 1: Performance metrics and 3D DSC of all models, per OAR (highlighted = baseline, red = best score per metric).

	2D Metrics						3D DSC per Organ				
Model	**AHD**	**ASSD**	**Dice**	**Jaccard**	**Precision**	**Recall**	**Esophagus**	**Heart**	**Trachea**	**Aorta**	**Mean**
Baseline-ENet	2.682	2.112	0.861	0.833	0.903	0.903	0.528	0.691	0.794	0.795	0.702
Best-ENet	2.742	2.166	0.891	0.862	0.914	0.919	0.538	0.718	0.823	0.808	0.722
UNet	2.771	2.264	0.907	0.879	0.936	0.925	0.611	0.643	0.835	0.865	0.738
UNetPlusPlus	2.291	1.803	0.910	0.887	0.951	0.919	0.625	0.710	0.859	0.887	0.770
DeepLabV3Plus	2.499	1.951	0.897	0.869	0.926	0.913	0.549	0.734	0.802	0.820	0.726

epochs and the Dice loss flatlines at 14 epochs in this specific case, but this behavior is observed across a variety of runs and across all classes, albeit at slightly different epochs. Given this behavior, we chose to use the combined loss for training all other models, except UNetPlusPlus, which performed better with cross-entropy loss.

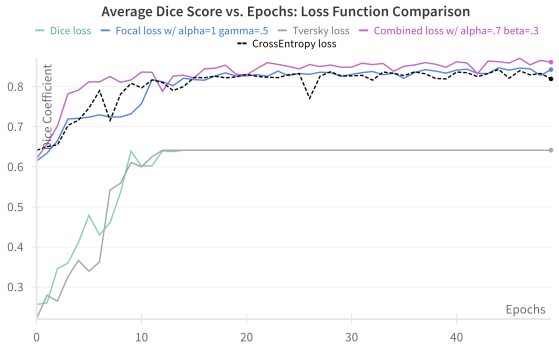


Fig. 2: Val DSC evolution during 50 epochs training of the best loss functions with the following best respective DSC: Focal: 0.847 - Combined: 0.871 - CE: 0.843 - Dice: 0.641

4.3. Comparison of models

Based on the previous results, our four final models have been run with data augmentation and Combined Loss, except UnetPlusPlus which was run with Cross-Entropy loss. The final scores for each 2D metric of our baseline and best-performing models can be found in Table 1. We can see that our fine-tuned Best-ENet achieves an increase of 3.5% in DSC compared to Baseline-ENet, highlighting the importance of stronger regularization methods for our given domain context. Furthermore, all the non-ENet models perform better than Best-ENet, suggesting that other architectures apart from ENet might be better suited for our thoracic segmentation problem. Notably, UNetPlusPlus is performing the best in all metrics except recall, where UNet is the highest achiever. This indicates that UNetPlusPlus provides slightly more conservative segmentations compared to UNet, but its segmentations are more accurate overall. Despite both UNet and DeepLabV3Plus showing good performance, their AHD and ASSD are significantly worse, suggesting the inconsistency in the shapes of their segmentations.

In addition to the 2D performance metrics discussed so far, Table 1 also shows the 3D Dice scores per OAR for our final models. That is, the evaluation is done on the whole patient at once instead of each slice. Again, we see that UNetPlusPlus performs the best out of all of the tested models for each organ. For all models, the esophagus achieves the lowest DSC, followed by the heart. Based on these results, we can assume that there's a correlation between 2D and 3D segmentation performance, at least for the dice metric.

A closer look at the average 2D Dice distribution across the 4 best performing models is offered in Figure 3. It can be observed that all distributions have a similar shape, with 2 density peaks, a larger one above the median at around 0.9 DSC and a smaller one below around 0.8 DSC. Additionally, while UNetPlusPlus has the distribution with the highest DSC density peak and the most concentrated spread, it also has the longest tail, indicating that its predictions are generally good and in the same high interval, but its outliers are more extreme.

The DSC distribution per organ between UnetPlusPlus and Best-ENet is presented in Figure 4. Both models have roughly the same shape for the tails, but UNetPlusPlus has its large densities and means situated on higher DSC scores than

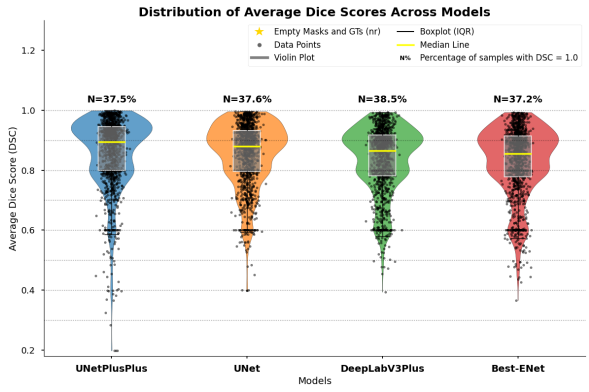


Fig. 3: Average validation 2D Dice distribution for all the models. Points with DSC 1.0 are excluded from the distribution and their counts are displayed on top for each model.

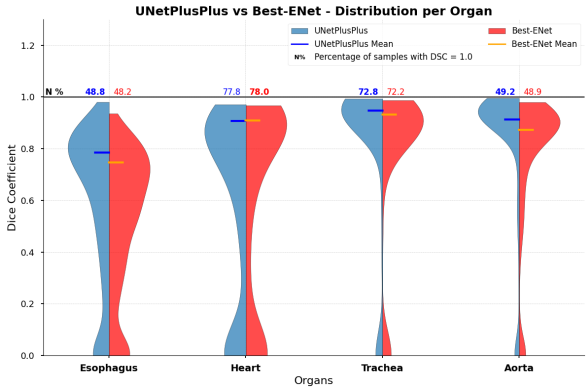


Fig. 4: Average 2D validation DSC distribution per organ - UNetPlusPlus VS Best-ENet. Percentages of points with a perfect DSC are displayed on top (highest ones in bold).

Best-ENet for every organ except the heart, where they are almost identical. UNetPlusPlus also exhibits a slight increase of 0.5% in points with a perfect DSC score.

Figure 5 provides more insight into the distribution of 2D Dice scores per organ for our best-performing UNetPlusPlus model. The esophagus has the lowest mean and more variance in the DSC distribution, suggesting unstable segmentations. The trachea and aorta show the highest performance both in mean and distribution of the scores, while the heart DSCs have the highest variance.

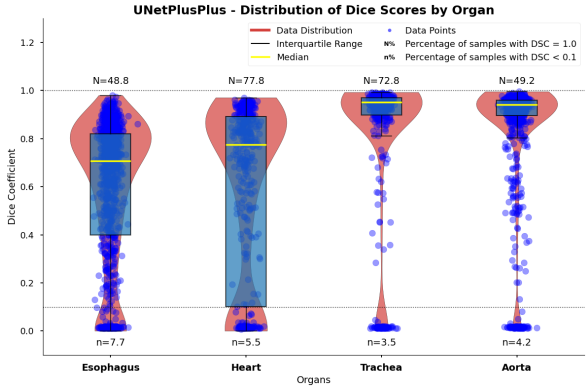


Fig. 5: Average 2D DSC distribution for UNetPlusPlus.

5. DISCUSSION

For this project, we had to work with relatively little data, concerning only 40 patients. This could result in a bias in the distribution which could become apparent when these models are evaluated on unseen data. Furthermore, due to limited computational resources all the reported results are obtained using only one random seed.

While all of our models work in 2D image space and a lot of our modeling choices are based on their 2D DSC performance, we recognize that this is an insufficient approach. Due to the nature of 2D metrics, 3D shape consistency is not taken into account, while this is important in a 3D clinical setting. While 3D DSC performance is used to analyze and compare our models, the use of 2D models and loss functions limits performance when evaluated in the full 3D setting of the dataset. We noticed a correlation between good 2D metrics performance and 3D DSC, however using a model that takes 3D shape into account during training should provide a better foundation for future approaches. These networks might even be able to get better 2D segmentations as a result. We would advice future research to look into the use of 3D architectures like VNet [12], which are likely to perform better in terms of intra-slice consistency and shape.

The observed performance using 2D metrics could explain the flattening of the Dice and Tversky loss. These measure the overlap between the ground-truth and predicted regions and in 2D medical imaging slices, this overlap can approach a Dice score close to 1, which leads to vanishing gradients. As a result, the model's performance stagnates.

Due to only having one patient showing the fixed position of the heart, we use a transformation based on only that patient. This was however not correct for all of the hearts, as some had different displacements. For future research, a more robust method of fixing the heart could be looked into in order to better position the segmentations.

While removing the empty ground truths reduced the Dice score of the model, we believe this to be due to the average being skewed higher with empty ground truths. Whenever both the prediction and the ground truth are empty, the Dice score returns 1, increasing the average. By removing the empty ground truths, we removed these scores from this average. Since the model still follows the same evolution curve as without the removal, the lower performance is an absolute decrease in performance. Future research should look into the viability of the removal of the empty ground truths from being taken into account for the average Dice score.

6. CONCLUSION

This paper aims to find the best CNN-like neural network for the segmentation of organs at risk in the thoracic region. Starting with a baseline implementation of ENet, we implemented four neural network architectures and compared their performance on the SegTHOR dataset. We found that out of the tested models, UNetPlusPlus has the highest performance overall both in 2D and 3D metrics, albeit with slightly lower recall. When looking at the Dice score distribution per organ for our best model, it becomes apparent that the esophagus and heart are the most difficult to segment, possibly due to under-representation in the dataset in terms of size and number of slice appearances, respectively, while the other organs perform better, possibly due to consistent shapes.

7. REFERENCES

[1] Z. Lambert, C. Petitjean, B. Dubray, and S. Ruan, "Segthor: Segmentation of thoracic organs at risk in ct images," 2019. [Online]. Available: https://arxiv.org/abs/1912.05950

[2] 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.

[3] A. Karimov, A. Razumov, R. Manbatchurina, K. Simonova, I. Donets, A. Vlasova, Y. Khramtsova, and K. Ushenin, "Comparison of unet, enet, and boxenet for segmentation of mast cells in scans of histological slices," in *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, 2019, pp. 0544–0547.

[4] A. Comelli, N. Dahiya, A. Stefano, F. Vernuccio, M. Portoghese, G. Cutaia, A. Bruno, G. Salvaggio, and A. Yezzi, "Deep learning-based methods for prostate segmentation in magnetic resonance imaging," *Applied Sciences*, vol. 11, no. 2, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/2/782

[5] P. V, V. R, and S. Chidambaranathan, "Polyp segmentation using unet and enet," in *2023 6th International Conference on Recent Trends in Advance Computing (ICRTAC)*, 2023, pp. 516–522.

[6] R. Zhao, B. Qian, X. Zhang, Y. Li, R. Wei, Y. Liu, and Y. Pan, "Rethinking dice loss for medical image segmentation," in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2020, pp. 851–860.

[7] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: a toolbox for intensity-based medical image registration," *IEEE transactions on medical imaging*, vol. 29, no. 1, pp. 196–205, 2009.

[8] D. P. Shamonin, E. E. Bron, B. P. Lelieveldt, M. Smits, S. Klein, M. Staring, and A. D. N. Initiative, "Fast parallel image registration on cpu and gpu for diagnostic classification of alzheimer's disease," *Frontiers in neuroinformatics*, vol. 7, p. 50, 2014.

[9] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka *et al.*, "3d slicer as an image computing platform for the quantitative imaging network," *Magnetic resonance imaging*, vol. 30, no. 9, pp. 1323–1341, 2012.

[10] M. I. Khalil, S. Tehsin, M. Humayun, N. Jhanjhi, and M. A. AlZain, "Multi-scale network for thoracic organs segmentation." *Computers, Materials & Continua*, vol. 70, no. 2, 2022.

[11] S. Liu, Y. Li, Q.-w. Chai, and W. Zheng, "Region-scalable fitting-assisted medical image segmentation with noisy labels," *Expert Systems with Applications*, vol. 238, p. 121926, 2024.

[12] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," 2016. [Online]. Available: https://arxiv.org/abs/1606.04797

[13] C. H. Sudre, W. Li, T. Vercauteren, S. 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*, M. J. Cardoso, T. Arbel, G. Carneiro, T. Syeda-Mahmood, J. M. R. Tavares, M. Moradi, A. Bradley, H. Greenspan, J. P. Papa, A. Madabhushi, J. C. Nascimento, J. S. Cardoso, V. Belagiannis, and Z. Lu, Eds. Cham: Springer International Publishing, 2017, pp. 240–248.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2018. [Online]. Available: https://arxiv.org/abs/1708.02002

[15] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," in *Machine Learning in Medical Imaging*, Q. Wang, Y. Shi, H.-I. Suk, and K. Suzuki, Eds. Cham: Springer International Publishing, 2017, pp. 379–387.

[16] M. E. Rayed, S. S. Islam, S. I. Niha, J. R. Jim, M. M. Kabir, and M. Mridha, "Deep learning for medical image segmentation: State-of-the-art advancements and challenges," *Informatics in Medicine Unlocked*, vol. 47, p. 101504, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914824000601

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[18] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," 2018. [Online]. Available: https://arxiv.org/abs/1807.10165

[19] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018. [Online]. Available: https://arxiv.org/abs/1802.02611

[20] V. Yeghiazaryan and I. Voiculescu, "Family of boundary overlap metrics for the evaluation of medical image segmentation," *Journal of Medical Imaging*, vol. 5, no. 1, p. 015006, 2018.

[21] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Research Notes*, vol. 15, no. 1, p. 210, 2022. [Online]. Available: https://doi.org/10.1186/s13104-022-06096-y

[22] N. Huynh, "Understanding evaluation metrics in medical image segmentation," https://medium.com/@nghihuynh_37300/understanding-evaluation-metrics-in-medical-image-segmentation-d289a373a3f, 2023, accessed: 2024-10-23.

[23] 3D Slicer Community, "3d slicer: Open source platform for biomedical research," https://www.slicer.org/, accessed: 2024-10-23.

A. CLASS IMBALANCE

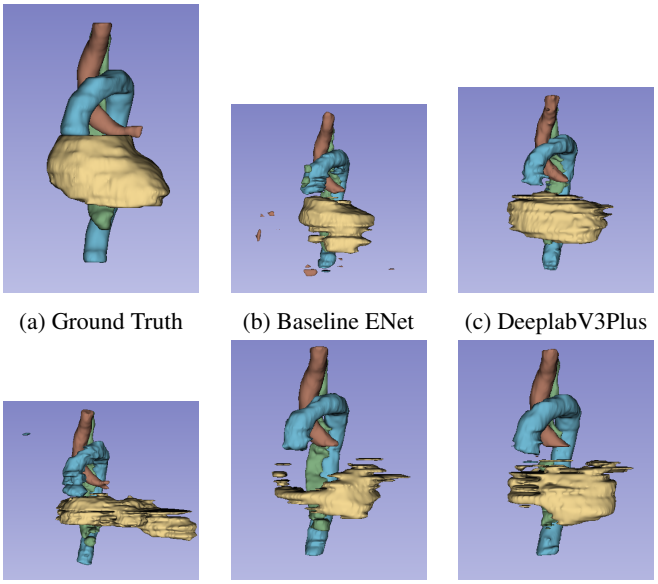
Table 2: The amount of slices present in each set, showing a large difference between the amount of slices per organ

	Train	Validation
Esophagus	2942	978
Heart	1203	396
Trachea	1495	492
Aorta	2797	957

Table 3: The average ratio of the area of each organ in the slices, showing that some organs are significantly larger than others

	Train	Validation
Esophagus	0.0581%	0.0482%
Heart	0.9423%	0.9893%
Trachea	0.0431%	0.0414%
Aorta	0.2468%	0.2251%

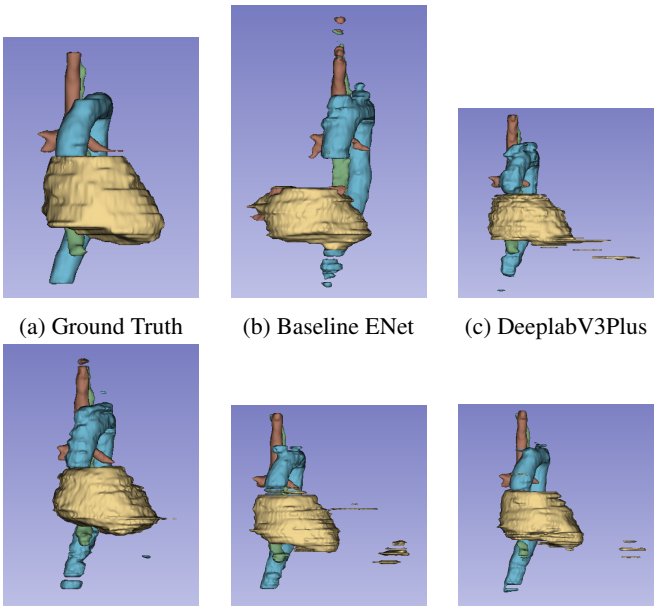
B. VISUAL RESULTS



(a) Ground Truth (b) Baseline ENet (c) DeeplabV3Plus

(d) Best ENet (e) UNet (f) UNetPlusPlus

Fig. 6: The segmentations for patient 01



(a) Ground Truth (b) Baseline ENet (c) DeeplabV3Plus

(d) Best ENet (e) UNet (f) UNetPlusPlus

Fig. 7: The segmentations for patient 13

C. PROJECT CONTRIBUTION

Task	Team Member	Description
Data Preprocessing + Insights	Thomas	Implement data augmentation techniques and prepare/inspect dataset
Model Architecture	Max + Alin	Enhance ENet & Implement Other Architectures, analysis of results
Loss Function	Leon	Implement and test Dice Loss, Focal Loss, Tversky Loss and combined loss
Metrics	Anesa + Thomas	Add new metrics for image segmentation