Project Report: Aorta Segmentation

Vijay Abhinav Telukunta

University of Florida, Gainesville, FL

Abstract. Aorta segmentation is a critical task in medical imaging, enabling accurate analysis of the aorta's structure and pathology. The integration of automatic aorta segmentation into clinical workflows offers significant benefits. It improves diagnostic accuracy, assists in surgical planning, and facilitates longitudinal studies of cardiovascular diseases. Automated segmentation of the aorta using deep learning methods has proven to be highly beneficial, addressing challenges such as time-consuming manual annotation, inter-observer variability, and scalability for large datasets. This project explores the performance of multiple state-of-the-art deep learning architectures, including UNet, VNet, SwinUNETR and CIS-UNet for aorta segmentation tasks on 3D medical imaging data. Quantitative evaluation metrics including Dice Similarity Coefficient, Hausdorff Distance and Surface Distance were employed to assess the segmentation performance. Qualitative results are also visualized. Among the models tested, CIS-UNet demonstrated superior performance in both quality and quantity.

1 Introduction

The segmentation of medical images is a critical task in medical image analysis, enabling precise delineation of anatomical structures and aiding in clinical diagnosis, treatment planning, and research. Among the various anatomical regions of interest, the aorta-a major artery in the cardiovascular system—holds significant importance due to its role in transporting oxygenated blood from the heart to the rest of the body. Accurate segmentation of the aorta in 3D medical imaging, such as computed tomography (CT) or magnetic resonance imaging (MRI), is essential for the assessment of conditions like aneurysms, dissections, and other vascular diseases.

Recent advancements in deep learning have revolutionized the field of medical image segmentation, providing powerful tools for automating the extraction of intricate structures from volumetric images. Traditional segmentation methods, which rely on manual delineation or heuristic-based approaches, are labor-intensive, prone to inter and intra-observer variability, and often fail to generalize across diverse datasets. In contrast, deep learning models, particularly UNET, VNET and transformer based models have demonstrated exceptional performance in capturing complex spatial and contextual information, making them well-suited for 3D segmentation tasks.

In this project, I aim to train and evaluate deep learning models for the multi-class 3D segmentation of the aorta using volumetric medical images. The task

involves distinguishing between multiple regions or branches of the aorta, such as the Innominate Artery, Right Subclavian Artery etc. each of which may present unique morphological and pathological characteristics. The key objectives of this project are as follows:

1. To preprocess 3D volumetric data effectively, ensuring uniformity in image dimensions, intensity normalization, and augmentation to enhance model robustness.

2. To implement and compare the performance of multiple deep learning architectures, assessing their accuracy, robustness, and efficiency in segmenting the aorta.

2 Related Work

Advancements in medical image segmentation have been greatly propelled by the emergence of deep learning techniques, particularly Convolutional Neural Networks (CNNs) and Transformer-based models.

2.1 CNN-Based Segmentation Models

U-Net and its variants have become foundational in medical image segmentation tasks due to their efficient design, which combines a downsampling path for feature extraction and an upsampling path for precise localization [5].V-Net [4] extends the U-Net architecture to volumetric data by employing 3D convolutions, making it particularly suited for tasks involving CT and MRI scans. Its encoder-decoder structure processes spatial relationships in three dimensions, enabling detailed volumetric segmentation of anatomical structures.

2.2 Transformer-Based Segmentation Models

Transformers, initially designed for natural language processing [7], have been successfully adapted to computer vision tasks. Vision Transformers (ViT) [1] have demonstrated the capability to model global relationships in image data, making them particularly suitable for complex segmentation problems. The Swin Transformer [3] introduced a window-based self-attention mechanism for multi-scale feature representation, significantly improving segmentation outcomes. SwinUNETR [6] incorporates a Swin Transformer encoder with a U-Net decoder. CIS-UNet [2], adapts the Swin Transformer with a novel Context-aware Shifted Window Self-Attention (CSW-SA) mechanism for multi-class segmentation.

3 Dataset

The AortaSeg24 dataset [2] is an essential tool for progressing multi-class segmentation in computed tomography angiography (CTA). It contains 100 annotated CTA volumes with precise labeling of the aorta, allowing for the differentiation of aortic branches and regions. This level of detail overcomes the

limitations of conventional binary segmentation techniques. Accurate segmentation, illustrated in Figure 1, is crucial for assessing the aorta's volume, diameter, and overall morphology.

In this project, the dataset included 50 3D aorta volumes stored in the MHA format. It is especially important for managing acute uncomplicated type B aortic dissection (auTBAD), a critical condition that demands detailed anatomical analysis. By aiding the development of automated segmentation algorithms, AortaSeg24 contributes to quicker and more accurate diagnoses, enhancing the precision of surgical planning. This initiative promotes collaboration across disciplines, advancing machine learning in medical imaging and improving outcomes for patients with complex aortic conditions.

This dataset was specially designed for aorta segmentation tasks and included detailed annotations for 24 distinct classes, each representing different anatomical regions of the aorta. The volumetric format of the data enabled a thorough assessment of the models' capabilities in segmenting intricate structures across multiple dimensions. Furthermore, the dataset's standardized format ensured compatibility with contemporary medical imaging tools, streamlining preprocessing and analysis.

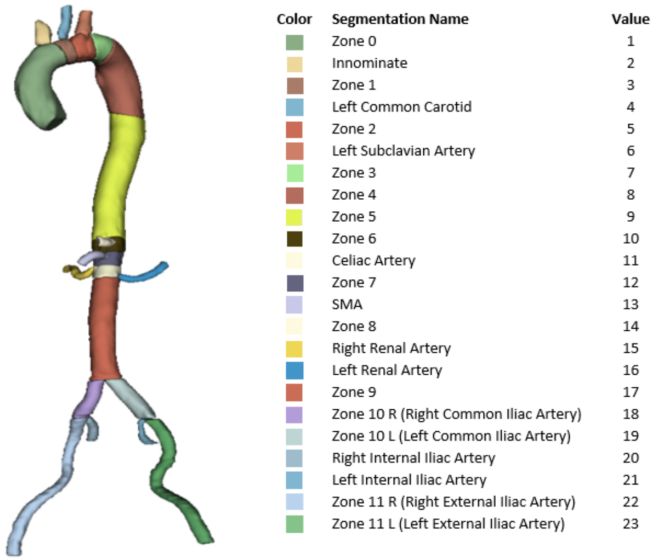

Fig. 1: Example of CTA input and corresponding multi-class segmentation labels

4 Method

4.1 Data Preprocessing

To prepare the 3D image volumes for training and validation in the aorta segmentation task, the following preprocessing steps were applied:

For Training Data, Loading: Images and masks were loaded in channel-first format. Intensity Normalization: Intensity values were scaled to the range [0, 1] with clipping. Foreground Cropping: Non-zero regions in the images were cropped to focus on the region of interest. Orientation Alignment: Data were reoriented to the RAS anatomical axis. Resampling: Data were resampled to a uniform voxel spacing of (2, 2, 2) mm^3 using bilinear interpolation for images and nearest-neighbor for masks. Random Cropping: Sub-volumes of size (128, 128, 128) were extracted, balancing positive and negative samples. Augmentations: Random flipping (10% probability per axis) and random 90° rotations were applied to improve model generalization.

Validation data underwent the same preprocessing, excluding augmentations, to preserve the anatomical structure for evaluation. These steps ensured consistent data formatting, enhanced model generalization, and focused learning on the aorta region. A total of 50 3D image volumes were used, with 40 volumes allocated for training and 10 volumes for validation.

4.2 Networks

To perform multi-class 3D segmentation of the aorta, the following deep learning architectures were utilized:

UNet: A classic encoder-decoder architecture with skip connections that enables efficient feature extraction and precise localization, suitable for medical image segmentation.

VNet: A volumetric fully convolutional network specifically designed for 3D medical image segmentation. It incorporates residual connections and an end-to-end training approach for better convergence on 3D volumes.

Swin UNETR: A transformer-based model combining the Swin Transformer for global self-attention with a U-Net-like architecture. This design enhances the network's ability to capture long-range dependencies in 3D medical images.

CIS UNet: It introduces a novel Context-aware Shifted Window Self-Attention (CSW-SA) block that enhances feature representation by capturing long-range dependencies between pixels.

All networks were implemented using the MONAI library, except CIS UNet, which was imported directly from its GitHub repository. These diverse architectures allowed for a comprehensive evaluation of model performance on the aorta segmentation task.

4.3 Loss Function and Optimizer

For training the segmentation models, DiceCELoss was used as the loss function. This is a weighted combination of the Dice Loss and Cross Entropy Loss,

effectively balancing region overlap and pixel-wise classification accuracy. Dice Loss emphasizes the overlap between predicted and ground truth regions, while Cross Entropy Loss ensures robust handling of class imbalance in multi-class segmentation.

Adam optimizer was employed to optimize the network parameters, leveraging adaptive learning rates for efficient convergence during training. The learning rate was set to 10^{-4} ensuring stable and consistent updates to the model weights.

4.4 Implementation details

All models, including UNET, VNET, SWINUNETR, and CISUNET, were trained for a total of 500 epochs. Validation was performed after every 5 epochs to monitor the model's performance on unseen data. During training, the model checkpoint was saved only if the validation Dice score improved compared to the previously recorded best Dice score. This approach ensured that the best-performing model on the validation set was retained for further evaluation.

For the training of all models, utilized an NVIDIA A100 GPU, a high-performance accelerator designed for AI workloads. The A100 offers significant computational power, especially for deep learning tasks, with its ability to handle large-scale data and complex models. This GPU was essential for accelerating the training process, reducing the time required to reach optimal model performance.

4.5 Evaluation Metrics

The Dice coefficient is a similarity measure that quantifies the overlap between two sets. It is commonly used for evaluating segmentation tasks, where the goal is to measure the similarity between the predicted segmentation and the ground truth. A Dice coefficient value close to 1 indicates a high similarity between the predicted and ground truth segmentations, while a value close to 0 indicates poor similarity.

Surface distance measures the geometric difference between the surfaces of the predicted and ground truth segmentations. This metric is particularly useful when you want to evaluate how well the boundaries of the predicted segmentation match the true boundaries. It is commonly used in medical image segmentation tasks to assess the accuracy of the surface prediction, which is important for applications like aorta segmentation.

The Hausdorff distance measures the greatest distance between a point in one segmented region (e.g., the predicted segmentation) and the closest point in the other region (e.g., the ground truth segmentation). It is a metric used to assess the worst-case deviation between the boundaries of two segmented objects. A Hausdorff distance value closer to 0 indicates a smaller discrepancy between the predicted and ground truth boundaries, signifying better alignment.

For the evaluation of segmentation performance, including the Dice coefficient (DC), Surface Distance (SD), and Hausdorff Distance (HD), I employed a two-step averaging process. First, for each individual image volume, I computed

the average metric across all 24 classes, which represents the mean value of Dice, Surface Distance, or Hausdorff Distance for that specific image. Then, the second mean was calculated by averaging the results of all image volumes in the validation set. This approach, referred to as the "mean of means," provides a robust evaluation of the models' performance across both individual classes and the entire dataset, ensuring that the overall segmentation quality is captured effectively.

5 Results

5.1 Quantitative Results

Model	MDC	MSD	MHD
U-Net	0.644	4.333	**17.256**
V-Net	0.674	**3.983**	18.931
Swin UNETR	0.659	7.832	35.449
CIS UNET	**0.695**	4.508	23.286

Table 1: Comparison of Mean Dice Coefficient(MDC), Mean Surface Distance (MSD) and Mean Hausdorff Distance(MHD) for Different Models

As per the results in Table 1, CIS UNET achieves the highest MDC, which suggests it has the best overall segmentation performance in terms of overlap with the ground truth. However, its MSD and MHD are slightly higher than V-Net and U-Net. V-Net offers a good balance with relatively high MDC and low MSD, though its MHD is slightly higher than UNET. U-Net, while not the best in terms of MDC, still offers competitive performance with the lowest MHD.

5.2 Qualitative Results

The qualitative results presented in the Figure. 2 illustrate the segmentation performance of different models in capturing the complex anatomy of the aorta and its branches. The ground truth clearly shows the aortic arch, descending aorta, and key branch vessels segmented with precise boundaries and minimal discontinuities. Among the models, CIS UNET achieves the closest visual match to the ground truth, accurately delineating the aortic regions and maintaining smooth, continuous segmentations of the branches. It successfully captures the finer structures, such as smaller branch vessels, without introducing noticeable artifacts.

In contrast, U-Net and V-Net exhibit clear limitations in accurately segmenting the aorta, particularly in capturing the smaller branches and maintaining consistent boundary alignment. These models tend to over-segment or leave gaps in critical regions, particularly in the aortic arch and descending aorta. Swin UNETR, while outperforming U-Net struggles with finer branch structures , which are better captured by CIS UNET.

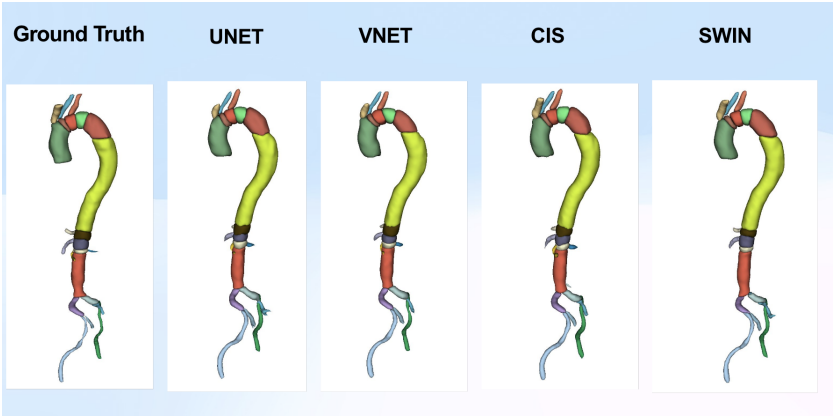

Fig. 2: Qualitative results.

6 Conclusion

This project focuses on training and evaluating various deep learning models for automated segmentation of aorta in image volumes. CIS-UNET stands out in achieving higher dice coefficient accurately segmenting the aortic regions. The project demonstrates the potential of deep learning-based segmentation models in improving medical image analysis and aiding clinical decision-making. Additionally, the collaboration between machine learning, medical imaging, and clinical applications could drive innovations that improve diagnostic accuracy and treatment outcomes for patients with complex aortic conditions. Future work could focus on refining segmentation accuracy, exploring advanced architectures, and incorporating additional data modalities to enhance model performance further.

References

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), https://arxiv.org/abs/2010.11929
2. Imran, M., Krebs, J.R., Gopu, V.R.R., Fazzone, B., Sivaraman, V.B., Kumar, A., Viscardi, C., Heithaus, R.E., Shickel, B., Zhou, Y., Cooper, M.A., Shao, W.: Cis-unet: Multi-class segmentation of the aorta in computed tomography angiography via context-aware shifted window self-attention (2024), https://arxiv.org/abs/2401.13049
3. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows (2021), https://arxiv.org/abs/2103.14030
4. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation (2016), https://arxiv.org/abs/1606.04797
5. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention pp. 234–241 (2015)
6. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B.A., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 20698–20708 (2021), https://api.semanticscholar.org/CorpusID:244715046
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6000–6010. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)