Aalto University
School of Electrical
Engineering

Master's Programme in ICT Innovation

Investigating Foundation Models in Medical Imaging

Jayesh Vasudeva

Master's Thesis
2024

Aalto University
School of Electrical Engineering

Author Jayesh Vasudeva

Title Investigating Foundation Models in Medical Imaging

Degree programme ICT Innovation

Major ICT Innovation

Supervisor Prof. Quan Zhou

Advisors Prof. Elena Troubitsyna, Prof. Pawel Herman

Collaborative partner Msc. Heikki Väänänen, AIATELLA Oy

Date November 17, 2024 **Number of pages** 78 **Language** English

Abstract

This thesis studies the efficacy of foundation models in medical image segmentation, focusing on CT scans amongst many modalities of medical imaging for the Aortaseg24 and HaN-Seg datasets. We implemented and fine-tuned six foundation models and a baseline Vanilla U-Net, evaluating their performance, adaptability, and computational efficiency. From our experiments we observe that foundation models consistently outperform the Vanilla U-Net baseline. STU-Net emerged as the top performer, achieving Dice scores of 0.76 for Aortaseg24 and 0.70 for HaN-Seg, significantly surpassing the baseline's 0.37 and 0.36 respectively. SwinMM followed closely, showing particularly strong performance on the more complex HaN-Seg task. Class-wise analysis (as Aortaseg24 and HaN-seg are multi-class segmentation datasets) revealed that foundation models excel in segmenting larger, more distinct anatomical structures, with STU-Net achieving Dice scores above 0.90 for numerous classes in Aortaseg24. However, all models struggled with smaller, more intricate structures, particularly in the HaN-Seg task. Computational efficiency varied widely among models. DAE processed the highest number of samples per second (2.99 ± 0.55), while SuPreM demonstrated the fastest convergence, requiring only 6000 iterations for Aortaseg24 and 4500 for HaN-Seg. Conversely, SPAD-Nets required the most iterations (30120 for Aortaseg24), highlighting significant differences in learning efficiency. The study also revealed task-specific performance variability. Models like SwinMM and MIS-FM showed better relative performance on HaN-Seg compared to Aortaseg24, highlighting that pre-training strategies and architectural designs influence a model's suitability for specific anatomical regions. Our findings highlight the potential of foundation models to progress medical image segmentation while also highlighting challenges in balancing performance with computational efficiency. The observed task-specific variability emphasizes the need for careful model selection based on the target application. This comprehensive analysis provides crucial insights for future developments in medical image segmentation, emphasizing the need for robust, versatile architectures capable of handling diverse and complex anatomical regions.

Keywords Medical Image Segmentation, Foundation Models, CT Scans, Deep Learning, Transfer Learning, Computational Efficiency, Aortaseg24, HaN-Seg

Preface

I would like to thank my colleagues at AIATELLA for their unbridled support and sparking the interest in me to pursue Deep Learning for Medical Imaging. I would also like to thank Prof. Pawel Herman without whom, I would be buried in a corpus of experiment, thank you for all your insights Professor.

I also extend my regards to Prof. Elena Elena Troubitsyna as my examiner and Heikki Väänänen for collaborating with AIATELLA as an advisor and helping me with the project.

Finally, I am grateful for all the love and support from my family and friends throughout the entire journey.

Helsinki, Finland, October 2024

Jayesh S Vasudeva

Contents

List of Figures

List of Tables

List of acronyms and abbreviations

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CMCL	Cross-modal contrastive learning
CNN	Convolutional Neural Network
CT	Computed Tomography
CTA	Computed Tomography Angiography
DA	Degree of anisotropy
DAE	Disruptive Autoencoders
GPT	Generative Pretrained Transformer
HU	Hounsfield Units
MIS-FM	Medical Image Segmentation using Foundation Models
MRI	Magnetic Resonance Imaging
OAR	Organs at Risk
PCT	Pixel-color transforms
PET	Positron Emission Tomography
RoPE	Rotary Positional Embeddings
SPAD-Nets	Spatially Adaptive Networks
SPAD-VT	Spatially Adaptive Visual Tokenizer
SPECT	Single Photon Emission Computed Tomography
STU-Net	Scalable and Transferable U-Net
SuPreM	Supervised Pre-trained Models
SwinMM	Masked Multi-view with Swin Transformers
VF	Volume Fusion

Chapter 1

Introduction

Medical imaging has revolutionized healthcare by enabling non-invasive visualization and analysis of internal anatomical structures. The exponential growth in medical imaging data, combined with advances in computational capabilities, has created unprecedented opportunities for improving patient care through automated analysis. The global medical imaging market is projected to reach $36.4 billion by 2024, growing at a CAGR of 6.9% from 2019 to 2024*. However, the complexity and variability of anatomical structures, coupled with the critical nature of medical decisions, make automated medical image analysis a particularly challenging domain

Foundation models, as defined by Bommasani et al. [1], are "models trained on broad data at scale that can be adapted to a wide range of downstream tasks." These models have shown remarkable success in natural language processing and computer vision tasks, prompting researchers to explore their potential in medical imaging. These models, pre-trained on vast datasets and adaptable to various downstream tasks, offer promising solutions for medical image analysis. However, their application to medical imaging introduces unique challenges that extend beyond traditional computer vision problems.

The segmentation of anatomical structures in CT scans presents unique challenges due to the complexity and variability of human anatomy. As illustrated in Figure 1.1, the distribution of different anatomical structures within a typical CT scan is highly imbalanced. For instance, in the AortaSeg24 dataset[2], we observe that Class 5 occupies 26.13% of the total segmented volume, while Class 1 accounts for merely 0.11%. This imbalance underscores

*Source: https://www.marketsandmarkets.com/Market-Reports/medical-imaging-market-411.html

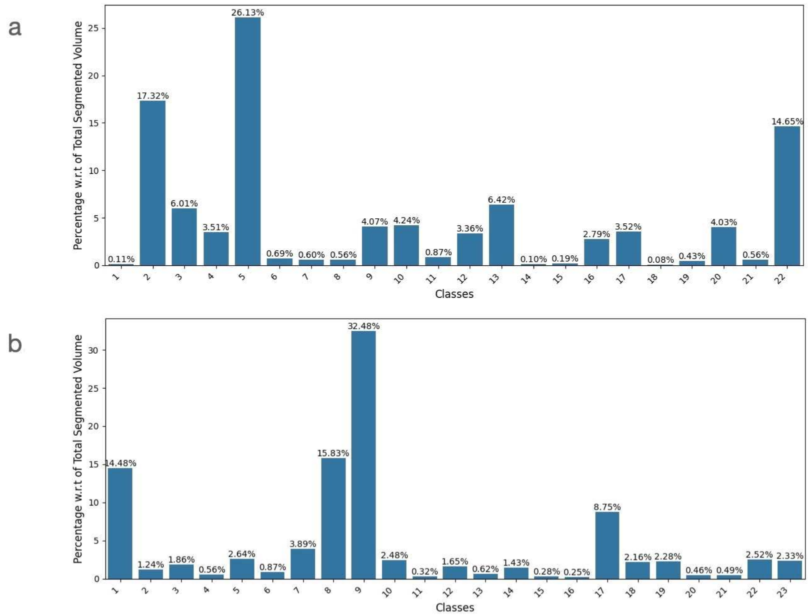

Figure 1.1: Distribution of classes within the non-zero volume of masks for (a) Aortaseg24 and (b) HaN-Seg

the need for models capable of handling such disparities effectively. The class imbalance not only sparks a statistical curiosity but also reflects the inherent complexity of human anatomy. Smaller structures, while occupying less volume, often carry uncertain, but significant clinical importance. For instance, in the context of aortic segmentation, accurately delineating smaller branches can be critical for planning endovascular procedures. The application of foundation models to this domain is particularly intriguing. These models, pre-trained on vast and diverse datasets, have the potential to capture intricate patterns and relationships that might elude traditional segmentation approaches. However, their effectiveness in handling the specific challenges of medical imaging remains an open question.

1.1 Problem Formulation

Recent approaches have taken diverse paths to address these challenges some have focused on specialized architectures [3], data augmentation techniques [4], and transfer learning approaches [5]. In particular Disruptive Autoencoders (DAE) introduced innovative self-supervised learning strategies through volume fusion and local masking[6], while Medical

Image Segmentation using Foundation Models (MIS-FM) demonstrated the potential of large-scale pre-training on Computed Tomography (CT) scans[7]. Spatially Adaptive Networks (SPAD-Nets) tackled the challenge of spatial heterogeneity through adaptive architectures[8], and Masked Multi-view with Swin Transformers (SwinMM) explored multi-view learning with transformer architectures[9]. These seminal works, whose implementations are publicly available(Table 3.1), form the foundation of our comparative study. While these methods have shown promise, they often struggle with:

- Feature Generalization: Models trained on specific anatomical regions often fail to generalize effectively to other structures. Despite efforts to develop more versatile architectures [10], the transfer of learned features across different anatomical regions remains suboptimal.

- Computational Efficiency: Current approaches frequently trade performance for computational cost. While some studies have proposed lightweight architectures [11], they often sacrifice accuracy on complex structures.

- Class Imbalance: Traditional loss functions and training strategies struggle with highly imbalanced anatomical structures. Recent work on adaptive loss functions [12] has shown improvements, but remains insufficient for clinical applications.

The central question emerging from these challenges remains: How do foundation models with different architectures and pre-training protocols along with a baseline U-Net architecture compare in their segmentation performance across diverse anatomical structures?

Our study investigates this question through a systematic comparison of six foundation models and a baseline architecture on the AortaSeg24[2] and HaN-Seg[13] datasets. By evaluating these models across multiple dimensions - including segmentation accuracy, computational efficiency, and generalization capability - we aim to provide insights into the effectiveness of foundation model architectures in medical image segmentation.

Our investigation centers on quantifying the extent to which foundation models' sophisticated feature representations translate to improved segmentation performance across diverse anatomical structures compared to specialized architectures. We evaluate six state-of-the-art foundation models against a baseline architecture on two challenging datasets: AortaSeg24[2] and HaN-Seg[13]. This comparative framework allows us to:

- Assess the generalization capabilities of foundation models across different anatomical regions
- Quantify the relationship between model complexity and segmentation performance
- Evaluate the practical viability of foundation models in clinical settings
- Identify specific architectural elements that contribute to successful feature transfer in medical imaging

The insights gained from this investigation will inform the development of more effective medical image segmentation systems and contribute to our understanding of how foundation models can be best leveraged in healthcare applications.

1.2 Objectives and scope

This study aims to provide a systematic comparison of foundation models in medical image segmentation, with specific objectives structured across three key dimensions:

- Implementation and Fine-tuning: Deploy and fine-tune six state-of-the-art foundation models for medical image segmentation Establish consistent evaluation protocols across all models Develop standardized pre-processing pipelines for CT imaging data
- Comparative Analysis: Evaluate segmentation performance across multiple anatomical structures Assess computational efficiency and resource requirements Analyze convergence behavior and training dynamics
- Performance Trade-offs: Investigate the relationship between model complexity and segmentation accuracy Examine task-specific performance variations across different anatomical regions Assess the practical implications of computational demands in clinical settings

1.2.1 Scope Limitations

1.2.1.1 Imaging Modality Restriction

Our analysis focuses exclusively on CT scans, limiting the generalizability of findings to other imaging modalities such as MRI or ultrasound. The distinct characteristics of CT imaging, including standardized Hounsfield units and consistent intensity distributions, mean our findings may not translate directly to modalities with different imaging physics and intensity patterns.

1.2.1.2 Dataset Specificity

The study evaluates models on two specific datasets: AortaSeg24 and HaN-Seg. While these datasets present complex, multi-class segmentation tasks, they represent only a subset of possible anatomical regions and pathological conditions. This limitation affects the breadth of conclusions we can draw about model performance across different medical contexts.

1.2.1.3 Single Training Instance

Each model is trained once on a single train-validation split, a significant limitation that prevents assessment of obvious performance variability across different data splits and training runs. This approach, while computationally efficient, means we cannot perform statistical hypothesis testing or provide confidence intervals for our results.

1.2.1.4 Computational Resource Constraints

The substantial computational requirements of foundation models, particularly during training and fine-tuning, limit our ability to perform multiple training runs or explore extensive hyperparameter optimization. This constraint affects both the robustness of our findings and their applicability to resource-constrained clinical environments.

1.2.1.5 Interpretability Considerations

Our study does not address the critical aspect of model interpretability, a key consideration for clinical adoption. The ”black box” nature of foundation models, especially larger architectures, poses challenges for regulatory approval and clinical trust that remain unexamined in our analysis.

1.2.1.6 Critical Statistical Limitation

It is crucial to emphasize that the absence of multiple training runs and subsequent statistical evaluation of performance variability represents a fundamental limitation of this study. Without proper statistical hypothesis testing, we cannot make definitive claims about the superiority of any model or architecture. The performance differences we observe, while potentially indicative of genuine trends, cannot be validated for statistical significance. Consequently, our findings should be interpreted as preliminary guidelines rather than conclusive results. This limitation particularly affects:

- Model performance comparisons
- Claims about convergence rates
- Efficiency metrics
- Task-specific performance variations

Future work should address these limitations through rigorous statistical analysis involving:

- Multiple training runs per model
- Cross-validation across different data splits
- Proper statistical hypothesis testing
- Confidence intervals for all reported metrics

This study therefore serves as a structured exploration of foundation models in medical image segmentation, providing insights and directions for more rigorous statistical investigations, rather than definitive conclusions about model superiority or optimal architectural choices.

1.3 Thesis Outline

This thesis comprises five chapters: Chapter 1 establishes the context of medical image analysis, introduces foundation models, and presents our research objectives focusing on comparative analysis of feature representations in medical image segmentation. Chapter 2 examines medical imaging modalities, foundation models, and analyzes six seminal architectures: Scalable and Transferable U-Net (STU-Net)[14], Supervised Pre-trained Models (SuPreM)[15], DAE[6], SPAD-Nets[8], MIS-FM[7], and SwinMM[9]. Reviews their theoretical foundations and architectural innovations. Chapter 3 details experimental framework, dataset specifications, implementation strategies, and evaluation metrics. Presents baseline architecture and comparative methodology. Chapter 4 presents multi-dimensional analysis of model performance across segmentation accuracy, computational efficiency, and convergence behavior. Examines task-specific variations across anatomical structures. Chapter 5 highlights interesting trends observed in Chapter 4 and provides an broader perspective on industrial, ethical and environmental impact of the study. Chapter 6 synthesizes key findings while emphasizing their preliminary nature, discusses implications for clinical applications, and provides recommendations for future statistical validation.

Chapter 2

Background

2.1 Medical Imaging Modalities

Medical imaging has become crucial in modern medicine, providing critical insights into the human body for point of care, diagnosis, treatment, and monitoring of various conditions[16]. This field contains a huge variety of tools and techniques, each with unique technologies and applications. The primary goal is to create visual representations of the interior of a patient's body for clinical analysis and medical intervention. It plays a vital role in improving patient outcomes by enabling non-invasive examination, thereby reducing the need for invasive methods and allowing for early detection of diseases and reduced costs.

2.1.1 X-ray

X-ray imaging, one of the oldest methods to capture osseous structures, primarily serves as a qualitative/point of care modality. Utilizing high-energy electromagnetic radiation, X-rays produce detailed images of the body's interior, making them particularly effective for visualizing bony structures and identifying fractures, infections, and tumors. As the X-rays pass through the body, various tissues absorb X-rays in varying degrees, thus revealing an image.

X-rays are produced when kinetic electrons interact with a solid target and undergo rapid deceleration. The resultant loss of kinetic energy is emitted as X-ray photons. These photons are subsequently captured by a conversion layer that absorbs the X-ray photons and emits electrons. These electrons are then accelerated and directed onto a luminescent output phosphor.

In the resulting images, dense materials such as bones, which absorb more X-rays, appear white, whereas softer tissues, which absorb fewer X-rays, are depicted in shades of gray. The ability of X-ray imaging to reveal minute details, combined with the relative affordability and user-friendliness of the equipment, contributes to its widespread utilization. However, it is important to note that X-ray imaging involves exposure to ionizing radiation, which carries a slight risk of inducing cancer with repeated exposures[17].

2.1.2 Computed Tomography(CT)

Computed tomography (CT) scans, also known as CAT scans, integrate multiple X-ray images taken from various angles to generate cross-sectional images of the body. The intensity of X-rays traversing the body is attenuated according to the density of the encountered tissues, thereby measuring the line integral of tissue density. During the scanning process, the X-ray source and detector rotate around the subject, collecting a row of X-ray measurements for each rotation angle. In helical CT scanners, the subject is continuously moved through the plane of the rotating source and detector, creating a helical trajectory around the subject. In multi-row helical CT scanners, the current state-of-the-art technology, the signal is acquired simultaneously from up to four rows, optimizing the balance between acquisition time, z-resolution, and noise.

The acquired signal, known as a sinogram, is represented as a series of images with detected attenuation on the x-axis and rotation angle on the y-axis. Image reconstruction involves solving the inverse Radon transform on the image grid. The reconstructed image displays attenuation values in Hounsfield units (Hounsfield Units (HU)), with water conventionally set to 0 HU. Consequently, fat appears with negative values, connective tissue with low positive values, and calcium with high positive values. Modern clinical scanners provide in-plane resolution of 0.5 mm and a slice thickness of 1 mm for anatomical structures within a single breath-hold.[18]

These images can be further processed to create three-dimensional representations. CT scans provide more detailed information than standard X-rays, making them particularly useful for diagnosing complex bone fractures, tumors, internal bleeding, and cardiovascular diseases. The use of contrast agents can enhance the visibility of specific areas, such as blood vessels and organs. However, similar to X-rays, CT scans involve exposure to ionizing radiation, necessitating a careful assessment of the associated risks and benefits. Despite their widespread use in clinical practice, CT angiography has certain drawbacks, including the X-ray dose absorbed by the patient and the potential intolerance to the iodinated contrast agents used[17].

2.1.3 Magnetic Resonance Imaging (MRI)

Magnetic resonance imaging (Magnetic Resonance Imaging (MRI)) alike CT is a volumetric imaging modality; however, the underlying physical principles are fundamentally different. While CT uses high-energy photons and the interaction of photons with electrons in the atomic shell to generate contrast, MRI is based on the orientation of protons within a strong magnetic field. This orientation can be manipulated with resonant radiofrequency waves, and the return of the protons to their equilibrium state can be measured.

The relaxation time constants, termed T_1 and T_2, are highly tissue-dependent, allowing MRI to provide superior soft tissue contrast, far exceeding that of CT. Consequently, MRI images can reflect proton density, T_1 , T_2, or a combination of these parameters.

Localization of voxels from the decaying magnetization signal, or free induction decay (FID), is achieved using additional magnetic fields that act as gradients along the imaging axes. The signal is collected as lines of data in the frequency domain, known as k-space, and is subsequently transformed into the image domain using Fourier transform techniques. Various sequences of magnetic field and gradient applications are available, providing significant flexibility in imaging different anatomical structures[18].

Furthermore, modern MRI scanners necessitate a superconductive magnet with a liquid helium cooling infrastructure, extremely sensitive radio-frequency amplifiers, and a room fully shielded against electromagnetic interference. Consequently, MRI equipment is exceedingly expensive, with costs reaching several million dollars for the scanner hardware and correspondingly high recurring maintenance expenses. However, MRI scanners deliver images of exceptionally high diagnostic value and can be used to monitor certain physiological processes (e.g., water diffusion, blood oxygenation), thus partly overlapping with nuclear imaging modalities[17].

2.1.4 Ultrasound

Ultrasound imaging, or sonography, employs high-frequency sound waves to generate real-time images of the body's interior. This technique relies on the emission of ultrasounds from a piezoelectric transducer, which also serves as a receiver for waves reflected at the interfaces between tissues with differing acoustic impedances. The brightness modulation (B-mode) imaging modality produces morphological images that represent the echogenicity of tissues and their interfaces. Recent advancements in image processing techniques have enabled the combination of acquired 2D images into 3D volumes, resulting in 3D ultrasound imaging. Although the resolution of ultrasound images is still not on par with that of CT or MR angiography and is limited to superficial vessels, this method is advantageous for its ability to accurately depict vessel wall morphology and for its relative cost-effectiveness and flexibility[18].

Ultrasound imaging is extensively utilized in obstetrics to monitor fetal development, as well as in cardiology, urology, and musculoskeletal imaging. This non-invasive technique does not employ ionizing radiation and is generally considered safe for all patients, including pregnant women. It is particularly effective for examining soft tissues and fluid-filled structures, such as the gallbladder, kidneys, and blood vessels. The quality of ultrasound images, however, is highly dependent on the skill of the technician performing the scan[17].

2.1.5 Nuclear Imaging

Nuclear imaging shares similarities with X-ray and CT imaging through its use of radiation, but differs fundamentally in its approach. Instead of external radiation sources, nuclear imaging involves injecting radioactive compounds into the body. These compounds are typically attached to pharmacologically active substances, known as radiopharmaceuticals, which accumulate in specific areas, such as tumors. Using either projection techniques or volumetric computerized image reconstruction, the spatial distribution of the radiopharmaceuticals can be mapped. This allows for the visualization of metabolic processes, which can be crucial for diagnostic purposes[18].

2.1.5.1 Positron Emission Tomography (PET)

Positron Emission Tomography (PET) scans offer functional insights into tissues and organs, making them instrumental in detecting cancer, heart disease, and brain disorders. PET is frequently combined with CT or MRI, allowing for the simultaneous acquisition of both functional and anatomical information. Although the use of radioactive tracers entails radiation exposure, the levels are typically low and deemed safe for most patients[17].

2.1.5.2 Single Photon Emission Computed Tomography(SPECT)

Alike PET, single photon emission computed tomography (Single Photon Emission Computed Tomography (SPECT)) is a nuclear medicine imaging technique that offers functional insights into the body's organs and tissues. SPECT involves gamma rays emitted by a radioactive tracer injected into the bloodstream. It is frequently utilized to diagnose and monitor conditions such as heart disease, bone disorders, and certain infections. Combining SPECT with CT can improve the anatomical detail of the images. As with other nuclear medicine techniques, radiation exposure is minimal and meticulously regulated.[17].

2.2 Foundation Models

Foundation models represent a revolutionary paradigm in artificial intelligence. As defined by Bommasani et al., these are *"base models trained on large-scale data in a self-supervised or semi-supervised manner that can be adapted for several other downstream tasks"*[1]. The concept of foundation models is rooted in the success of large language models in natural language processing, such as Generative Pretrained Transformer (GPT)-3[19] and Bidirectional Encoder Representations from Transformers (BERT)[20].These models represent a shift from task-specific approaches towards more universal and adaptable frameworks for medical image analysis. However, their application in medical imaging introduces unique challenges and opportunities. The core idea is to develop a single, highly capable model that can be fine-tuned for various specific tasks, rather than creating separate models for each application.

One of the key characteristics of foundation models is their ability to learn from diverse and unstructured data sources, such as text, images, and videos[20]. By leveraging self-supervised learning techniques, such as masked language modeling and contrastive learning, foundation models can capture intricate patterns and relationships within the data without requiring manual labeling[21]. This allows foundation models to develop a deep understanding of the underlying structure and semantics of the data, enabling them to generate coherent and contextually relevant outputs.

Another important aspect of foundation models is their adaptability to various downstream tasks. Once a foundation model is trained on a large-scale dataset, it can be fine-tuned on specific tasks with relatively small amounts of labeled data[19]. This transfer learning capability enables foundation models to achieve state-of-the-art performance on a wide range of tasks, such as language translation, question answering, image classification, and more[22]. The adaptability of foundation models has significant implications for reducing the cost and effort required to develop specialized Artificial Intelligence (AI) systems.

In conclusion, foundation models learn from raw data without explicit supervision, allowing them to capture rich semantic information and generalize to new tasks with minimal fine-tuning[23]. Foundation models represent a significant advancement in artificial intelligence, offering a powerful and adaptable approach to learning from diverse and unstructured data. As research in this area continues to progress, foundation models have the potential to transform various domains and enable new applications of AI. However, it is crucial to carefully consider the ethical implications and develop responsible practices for the development and deployment of foundation models.

2.2.1 Architectures and Pre-training

One of the most prominent architectures used in foundation models for medical imaging is the Transformer. Transformers, originally introduced for natural language processing tasks, have been adapted to handle visual data, including medical images. The self-attention mechanism in Transformers allows the model to capture long-range dependencies and learn contextual relationships between different regions of an image. Dosovitskiy et al. proposed the Vision Transformer (ViT), which applies the Transformer architecture directly to image patches, demonstrating its effectiveness in learning visual representations[24].

Building upon the success of ViT, several variants have been developed specifically for medical imaging tasks. Matsoukas et al. introduced Medical Transformer (MedT), which incorporates domain-specific modifications to the ViT architecture, such as using 3D convolutions and integrating spatial priors, to better capture the characteristics of medical images[25]. Another notable architecture is the Swin Transformer, proposed by Liu et al., which employs a hierarchical design and shifted window attention to efficiently process high-resolution medical images while maintaining computational efficiency[26].

Pre-training strategies play a crucial role in the success of foundation models for medical imaging. Self-supervised learning has emerged as a powerful approach to learn meaningful representations from unlabeled data. Contrastive learning, a popular self-supervised technique, aims to learn representations that are invariant to certain transformations while being discriminative between different instances. Valanarasu et al. applied contrastive learning to medical images, using a combination of different modalities to allow the model learn diversities across modalities in a single framework[6]. Another method by Sowrirajan et al. applied contrastive learning to medical images, using a combination of global and local contrastive losses to capture both global structure and local details[27].

Another effective pre-training strategy is masked image modeling, inspired by the masked language modeling used in natural language processing. In this approach, a portion of the input image is masked, and the model is trained to reconstruct the missing regions based on the surrounding context. Zhou et al. proposed the Masked Autoencoders Are Scalable Vision Learners (MAE) framework, which achieves state-of-the-art performance on various medical image analysis tasks by pre-training on large-scale unlabeled datasets[28].

The pre-training of foundation models on diverse medical image datasets has shown significant benefits in terms of generalization and transferability. Azizi et al. demonstrated the effectiveness of pre-training on a large-scale chest X-ray dataset, showing improved performance on multiple downstream tasks, including disease classification and segmentation[29]. . Similarly, Matsoukas et al. pre-trained their Medical Transformer on a multi-modal dataset consisting of CT, MRI, and PET images, achieving strong results on various medical image analysis benchmarks [25].

However, there are challenges and considerations when applying foundation models to medical imaging. One major concern is the potential bias and fairness issues that may arise from the pre-training data. Medical image datasets often suffer from class imbalance, demographic biases, and variations in image acquisition protocols. Careful curation and stratification of the pre-training data, along with techniques like data augmentation and domain adaptation, can help mitigate these issues[30].

Another challenge is the interpretability and explainability of foundation models in medical imaging. As these models become more complex and capture intricate patterns, understanding their decision-making process becomes crucial for clinical adoption. Techniques such as attention visualization, saliency maps, and concept-based explanations have been explored to provide insights into the model's reasoning[31].

2.2.2 Adaptations and Fine-tuning

Foundation models have demonstrated remarkable capabilities in learning general-purpose representations from large-scale datasets. However, to effectively apply these models to specific medical imaging tasks, adaptation and fine-tuning techniques are crucial. These techniques allow the models to leverage the knowledge acquired during pre-training and tailor it to the unique characteristics and requirements of medical imaging applications.

One common approach for adapting foundation models to medical imaging tasks is transfer learning. Transfer learning involves initializing the weights of a target model with those learned from a pre-trained foundation model and then fine-tuning the model on a specific medical imaging dataset. This approach has been widely adopted in various medical imaging applications, such as disease diagnosis, segmentation, and anomaly detection.

Matsoukas et al. proposed the Medical Transformer (MedT), which adapts the Vision Transformer (ViT) architecture for medical image analysis[25]. They fine-tuned the pre-trained ViT model on a diverse set of medical imaging datasets, including CT, MRI, and PET images, and demonstrated improved performance on various downstream tasks compared to training from scratch. The fine-tuning process allowed the model to learn domain-specific features and capture the nuances of medical images.

Another notable work is the adaptation of the Segment Anything Model (SAM) for medical image segmentation. Hu et al. explored efficient techniques to adapt SAM to medical images, leveraging prompt engineering and fine-tuning strategies[32]. By carefully designing prompts and fine-tuning the model on a small set of annotated medical images, they achieved impressive segmentation results across different imaging modalities and anatomical structures. This showcases the potential of foundation models to accelerate the development of accurate segmentation models in medical imaging with limited labeled data.

Few-shot learning is another promising approach for adapting foundation models to medical imaging tasks with scarce labeled data. Azizi et al. demonstrated the effectiveness of few-shot learning using a pre-trained foundation model for skin lesion classification[29]. By fine-tuning the model on a small number of labeled examples from each class, they achieved performance comparable to dermatologists. This highlights the potential of foundation models to enable accurate diagnosis even in scenarios where obtaining large amounts of labeled medical data is challenging.

Domain adaptation techniques have also been explored to bridge the gap between the pre-training data and the target medical imaging domain. Deng et al. proposed a domain adaptation framework that leverages self-supervised learning and adversarial training to adapt foundation models to medical images[33]. By aligning the feature distributions of the pre-training and target domains, they improved the transferability of the foundation model to various medical imaging tasks, such as disease classification and segmentation.

Fine-tuning foundation models on multi-modal medical data has shown promising results in capturing complementary information and improving overall performance. Nguyen et al. introduced LVM-Med, a framework for learning large-scale self-supervised vision models for medical imaging using second-order graph matching[34]. By leveraging both visual and textual information from medical records, they fine-tuned the foundation model to learn rich representations that capture the relationships between images and associated clinical data. This multi-modal fine-tuning approach has the potential to enhance the interpretability and reliability of medical imaging models.

2.3 Seminal Works

In this section, we discuss a few seminal works that form the foundation for our thesis on foundation models in medical imaging.

2.3.1 Scalable and Transferable U-Net(STUNet)

STU-Net displays the effectiveness of scaling segmentation models to unprecedented sizes, with the largest variant containing 1.4 billion parameters[14]. This scale-up allows the model to capture more complex features and relationships in medical images, potentially leading to improved accuracy across a wide range of segmentation tasks. Secondly, STU-Net's architecture incorporates refinements to the popular nnU-Net framework[35], such as residual connections and improved upsampling techniques, which enhance both scalability and transferability[14]. These modifications enable the model to be easily adapted to various medical imaging tasks without significant architectural changes.

The model's pre-training on the TotalSegmentator dataset[36], which includes 104 anatomical structures from 1204 CT images, provides it with a broad understanding of human anatomy.This comprehensive pre-training contributes to STU-Net's strong transfer learning capabilities, allowing it to perform well on new tasks with minimal fine-tuning. Huang et al. demonstrate this transferability through extensive evaluations on 14 downstream datasets for direct inference and 3 datasets for fine-tuning, covering a diverse range of anatomical structures and imaging modalities. When comparing STU-Net to other foundation models, its specialization in volumetric medical image segmentation is a crucial consideration. Unlike general-purpose vision models, STU-Net is tailored to the unique challenges of 3D medical imaging, making it particularly relevant for healthcare applications. The model's ability to handle multiple organs and structures simultaneously also sets it apart from many existing medical segmentation models that focus on specific anatomical regions. Furthermore, STU-Net's performance improvements with increasing model size suggest that it may benefit from future scaling efforts, similar to trends observed in natural language processing and computer vision[14].

STU-Net's approach of using supervised pre-training on large, diverse dataset of annotated medical images represents a step towards developing true foundation models for medical imaging, This contrasts with some other approaches that rely on self-supervised learning or transfer from natural image domains [7, 6, 37]. The success of this strategy in producing highly transferable models could influence the direction of future research in medical AI. Additionally, Huang et al's exploration of the trade-offs between universal models (trained on all anatomical structures) and expert models (specialized for specific structures) provides valuable insights for designing flexible, multi-purpose medical imaging systems [14].

2.3.2 Disruptive Autoencoders(DAE)

DAE introduces a novel pre-training strategy called Volume Fusion (Volume Fusion (VF)), which creates a pseudo-segmentation task from unlabeled 3D medical images [6]. This approach ingeniously bridges the gap between self-supervised pre-training and downstream segmentation tasks, addressing a key limitation of previous methods. VF fuses several random patches from a foreground sub-volume to a background sub-volume based on predefined discrete fusion coefficients, creating a challenging yet relevant pretext task. One of the most innovative aspects of DAE is its ”local masking” strategy. Unlike traditional masked autoencoder approaches that mask entire tokens, DAE masks at the channel embedding dimension [6]. This subtle but crucial difference allows the model to preserve and learn from fine-grained local details, which are essential in medical imaging. By retaining some information for each token, DAE enables better reconstruction of small anatomical structures and lesions that might be entirely masked out in traditional approaches. DAE's multi-faceted disruption approach further enhances its learning capabilities. By combining local masking with classical low-level perturbations like noise addition and downsampling, DAE creates a rich, challenging pretext task [6]. This combination forces the model to learn robust features that are resilient to various types of image degradation, a valuable trait in real-world medical imaging scenarios where image quality can vary significantly. The introduction of a cross-modal contrastive learning (Cross-modal contrastive learning (CMCL)) component in DAE is particularly noteworthy. This allows the model to learn to differentiate between features from different imaging modalities while maintaining a unified framework [6]. In the context of medical imaging, where multi-modal analysis is often crucial, this capability is extremely valuable. It potentially allows for more effective transfer learning across different imaging modalities, a persistent challenge in medical AI. DAE was pre-trained on a large-scale dataset of over 10,000 3D volumes across multiple modalities including CT and various MRI types (T1, T2, FLAIR, T1ce) [6]. This diverse pre-training dataset allows the model to learn a rich set of features that are potentially applicable across a wide range of medical imaging tasks and modalities. The scale and diversity of this dataset contribute significantly to the model's generalization capabilities. One of the most compelling aspects of DAE is its performance in reconstruction tasks. The model demonstrates an ability to reconstruct fine details and structures that are often lost in other self-supervised learning approaches [6]. This capability suggests that DAE is learning a particularly rich and medically relevant feature representation, which is crucial for downstream tasks like segmentation and anomaly detection. DAE shows remarkable data efficiency in downstream tasks. It often achieves comparable or better performance than fully supervised models while using only a fraction of the labeled data for fine-tuning [6]. This efficiency is particularly critical in medical imaging, where labeled data is often scarce and expensive to obtain. The model's performance on downstream tasks is impressive, showing strong results across multiple segmentation benchmarks, particularly when fine-tuned on limited labeled data.

2.3.3 Spatially Adaptive Networks(SPADNets)

SPAD-Nets address a fundamental challenge in medical image analysis. At their core, these models address a fundamental challenge in medical imaging: the heterogeneity of spatial properties across different imaging modalities and acquisition protocols[8]. The spatially adaptive networks (SPAD-Nets) introduce a novel approach to handling this heterogeneity. Unlike traditional models that require task-specific architectures, SPAD-Nets dynamically adjust their structure based on the input image's properties. This is achieved through the innovative spatially adaptive convolution(SPAD-conv), which shares parameteres with a conventional 3D convolution but transforms them on-the-fly to adapt to the input's spatial characterisitics. For instance, when encountering an image with highly anisotropic voxel spacing (where the slice thickness is much larger than the in-plane spacing), SPAD-conv automatically adjusts its kernel size and stride along the depth dimension[8]. This adaptive behavior allows the model to maintain consistent performance across a wide range of image properties without the need for manual architecture adjustments or image resampling.

A particularly intriguing aspect of SPAD-Nets is their implementation of a generalized resampling technique. This approach introduces the concept of ”degree of anisotropy” (Degree of anisotropy (DA)), which quantifies how anisotropic an image is. The model uses this DA to determine how to adjust its downsampling and upsampling operations, ensuring that it processes images efficiently while preserving important spatial information[8]. This nuanced handling of spatial properties is crucial in medical imaging, where preserving the correct spatial relationships can be critical for accurate diagnosis and analysis.

The development of a spatially adaptive visual tokenizer (Spatially Adaptive Visual Tokenizer (SPAD-VT)) addresses the lack of suitable pre-trained tokenizers for medical images. SPAD-VT employs a ”soft token representation,” where tokens are represented as distributions rather than discrete entities. This approach, combined with a dual prior distribution regularization technique, helps mitigate common issues in visual tokenization such as codebook collapse and low codebook utilization[38]. The resulting tokenizer demonstrates remarkable versatility, capable of reconstructing images from diverse modalities and anatomical regions with high fidelity. By incorporating 55 public datasets spanning multiple modalities and anatomical regions, the model learns a rich, generalizable representation of medical imaging data. This comprehensive pre-training enables the model to perform well on a wide range of downstream tasks, even with limited task-specific training data. A particularly noteworthy aspect is SPAD-Nets's performance in few-shot learning scenarios. The experiments demonstrate that the pre-trained model can achieve better segmentation results using only 60% of the training data compared to a model trained from scratch on the full dataset[8]. This capability has significant implications for medical AI, where labeled data is often scarce and expensive to obtain. The extension of rotary position embedding (Rotary Positional Embeddings (RoPE)) to 3D medical images is another subtle but important innovation. By carefully designing an additive 3D RoPE, the model can effectively capture spatial relationships in 3D volumes while maintaining compatibility with pre-trained 2D weights. This thoughtful integration of positional information contributes to the model's ability to handle both 2D and 3D images effectively. When comparing SPAD-Nets to other foundation models, their unique ability to handle spatial heterogeneity stands out. SPAD-Nets adaptive architecture allows them to process a much wider range of medical imaging data without the need for specialized variants or extensive preprocessing[8].

2.3.4 Medical Image Segmentation using Foundation Models(MIS-FM)

MIS-FM represents a significant leap in developing foundation models for 3D medical image segmentation. At its core, the Volume Fusion (VF) strategy addresses a fundamental challenge in medical imaging: the scarcity of annotated data. Unlike previous self-supervised learning approaches that focus on image-level tasks or restoration, VF ingeniously formulates the pretraining task as a pseudo-segmentation problem. This alignment between pretraining and downstream tasks is crucial, as it allows the model to learn features that are directly relevant to segmentation[7].

A particularly intriguing aspect of VF is its ability to encourage multi-scale context learning. By randomly generating foreground patches at various scales and aspect ratios, and using soft fusion coefficients, VF forces the model to understand both local and global contexts[7]. This multi-scale learning is essential for medical imaging, where structures of interest can vary greatly in size and shape across different anatomical regions.

The Pixel-color transforms (PCT)-Net architecture introduces a novel approach to combining Convolutional Neural Network (CNN)s and Transformers for 3D medical image analysis. Its design is particularly clever in handling the anisotropic nature of many medical images. By using 2D convolutions at the highest resolution level and transitioning to 3D convolutions at lower resolutions, PCT-Net efficiently processes images with different in-plane and inter-plane spacings[7]. This approach addresses a common issue in 3D medical image analysis where naive 3D architectures can be computationally inefficient or struggle with anisotropic data.

The pyramid PCT module in PCT-Net is another innovative feature. By using parallel convolutional and transformer branches at multiple resolution levels, the model can simultaneously capture local details and long-range dependencies[7]. This is crucial in medical imaging, where both fine-grained tissue characteristics and overall anatomical context are important for accurate segmentation. With 110k CT scans covering various body parts and pathologies, the model has the potential to learn a rich, generalizable representation of human anatomy. This large-scale pretraining is particularly valuable given the diversity of the dataset, including images from different vendors, scanning protocols, and patient demographics. Such diversity in the pretraining data enhances the model's ability to generalize across different downstream tasks and datasets.

A crucial aspect of MIS-FM is its computational efficiency. Despite its large-scale pretraining and sophisticated architecture, the model is designed to be usable on common GPU devices[7]. This practicality is essential for widespread adoption in clinical and research settings, where access to high-end computing resources may be limited. The experimental results of MIS-FM are particularly impressive in their consistency across different anatomical regions and organs. The model's ability to outperform specialized state-of-the-art models on tasks ranging from brain stem segmentation to liver segmentation demonstrates its potential as a truly universal foundation model for medical image segmentatio[7]. This versatility could significantly streamline the development of AI-assisted diagnosis and treatment planning tools across various medical specialties. Moreover, the model's performance in low-data regimes is especially noteworthy. The ability to achieve high performance with only 60% of the training data compared to models trained from scratch highlights the power of the pretrained features[7]. This characteristic is particularly valuable in medical imaging, where acquiring large, annotated datasets for rare conditions or novel imaging protocols can be prohibitively expensive and time-consuming.

2.3.5 Multi-view with Swin Transformers(SwinMM)

The model employs a masked multi-view encoder that simultaneously processes multiple masked perspectives of 3D medical images during pretraining[9]. This novel approach allows the model to construct rich, multi-dimensional representations that integrate information from various angles, which is particularly crucial for the complex spatial relationships present in 3D medical data. By learning to reconcile these different views, SwinMM develops a more comprehensive understanding of anatomical structures than models trained on single views alone. The pretraining phase of SwinMM is notable for its use of four complementary proxy tasks: image reconstruction, rotation prediction, contrastive learning, and a novel mutual learning task[9]. This diverse set of tasks is carefully designed to encourage the model to learn a wide range of features. The image reconstruction and rotation tasks focus on local spatial relationships, while the contrastive learning promotes the extraction of global features. The mutual learning task, a unique contribution of SwinMM, enforces consistency across different views, teaching the model to recognize the same anatomical structures from various angles. This multi-faceted approach to pretraining results in a model with a rich, nuanced understanding of 3D medical imagery. For the fine-tuning phase, SwinMM introduces an innovative cross-view decoder featuring a cross-attention mechanism[9]. This decoder is designed to effectively aggregate and synthesize information from multiple views, allowing the model to make more informed segmentation decisions. The cross-attention mechanism enables the model to dynamically focus on the most relevant information from each view, potentially leading to more accurate segmentations, especially in complex or ambiguous cases. This multi-view integration during both pretraining and fine-tuning sets SwinMM apart from many other foundation models in medical imaging. One of the most impressive aspects of SwinMM is its data efficiency. In experiments, the model demonstrated superior performance with limited labeled data, outperforming competing methods when using only a fraction of the available training data[9]. For instance, with just 30% of labeled data, SwinMM achieved a Dice score of 78.91%, compared to 70.65% for Swin UNETR. This efficiency is crucial in the medical domain, where large-scale annotated datasets are often difficult and expensive to obtain. The model's ability to learn effectively from limited data could significantly reduce the annotation burden in clinical settings. SwinMM's versatility is another key strength. The model has shown strong performance across a variety of anatomical regions and imaging modalities, from cardiac MRI to abdominal CT scans[9]. This broad applicability suggests that SwinMM has potential as a universal foundation model for medical image segmentation, capable of adapting to diverse clinical scenarios with minimal additional training. Such versatility could streamline the deployment of AI-assisted segmentation tools in clinical practice, reducing the need for multiple specialized models. The scale of pretraining for SwinMM is noteworthy. The model was pretrained on over 5800 3D volumes from 8 public datasets, representing one of the largest pretraining efforts for 3D medical segmentation to date[9]. This extensive pretraining allows the model to learn from a diverse range of anatomical structures and imaging characteristics, contributing to its robust performance across different tasks. The ability to effectively leverage such large-scale, unlabeled data is a significant advantage in the medical imaging field, where labeled data is scarce but unlabeled scans are often abundant. In terms of performance, SwinMM sets new state-of-the-art benchmarks. On the WORD dataset, it achieved an average Dice score of 86.18%, surpassing both fully-supervised and self-supervised baselines[9]. Notably, it outperformed the previous best model, LCOVNet, by 0.36 percentage points in Dice score while also reducing the Hausdorff distance. These improvements, while seemingly small, can translate to significant clinical impact, potentially leading to more accurate diagnoses and treatment planning. An intriguing aspect of SwinMM is its extensibility to semi-supervised learning scenarios. The model's multi-view approach can be readily adapted to leverage both labeled and unlabeled data, further enhancing its data efficiency[9]. In semi-supervised experiments, SwinMM demonstrated a remarkable ability to improve performance with limited labeled data, outperforming single-view approaches by a significant margin. This capability could be particularly valuable in scenarios where obtaining large amounts of labeled data is challenging, such as for rare diseases or novel imaging protocols.

2.3.6 Supervised Pre-trained Models(SuPreM)

SuPreM leverages AbdomenAtlas 1.1, an unprecedented dataset comprising 9,262 CT volumes with per-voxel annotations of 25 anatomical structures and pseudo annotations of seven tumor types [15]. This dataset is not just large in scale, but also remarkably diverse, encompassing CT scans from 88 hospitals worldwide. This diversity is crucial for developing robust models that can generalize across different imaging protocols and patient populations. One of the most striking aspects of SuPreM is its extraordinary data efficiency. In a direct comparison, SuPreM trained on just 21 CT volumes (with 672 masks) for 40 GPU hours demonstrated transfer learning abilities comparable to self-supervised models trained on 5,050 CT volumes for 1,152 GPU hours [15]. This represents a 99.6% reduction in data requirements and a 96.5% reduction in computational costs. Such efficiency is game-changing in the medical field, where annotated data is often scarce and expensive to obtain. SuPreM's pre-training strategy is uniquely aligned with downstream segmentation tasks. Unlike self-supervised methods that rely on pretext tasks like masked image modeling or contrastive learning, SuPreM directly learns to segment anatomical structures [15]. This direct alignment results in learned features that are inherently more relevant to medical image segmentation tasks. The benefits of this approach are evident in the model's performance on novel classes: SuPreM outperforms self-supervised models in segmenting 63 anatomical structures that were not part of its pre-training set [15]. The architecture of SuPreM is another area of innovation. It incorporates both CNN and Transformer elements, allowing it to capture both local details and global contextual information effectively [15]. This hybrid approach is particularly well-suited to medical imaging, where both fine-grained features (like small lesions) and overall anatomical context are crucial for accurate diagnosis and segmentation. SuPreM's generalization capabilities are particularly noteworthy. In direct inference tests on external datasets from different countries and hospitals, SuPreM achieved performance comparable to or even exceeding models specifically trained on those datasets [15]. For instance, on the TotalSegmentator dataset (representing the Central European population), SuPreM achieved an average Dice score of 85.9%, compared to 84.4% for the dataset-specific model. This cross-domain performance is crucial in medical imaging, where data distribution can vary significantly across institutions and populations. In the realm of tumor detection and classification, SuPreM demonstrates remarkable capabilities. When fine-tuned for pancreatic tumor classification, it achieved a sensitivity of 86.1% and specificity of 95.4% for PDAC detection, surpassing average radiologist performance by 27.6% in sensitivity and 4.4% in specificity [15]. This level of performance on such a challenging task underscores the model's potential to significantly impact clinical practice. SuPreM also exhibits impressive few-shot learning capabilities. In experiments with limited data, SuPreM consistently outperformed self-supervised models, achieving higher accuracy with fewer labeled examples [15]. This characteristic is particularly valuable in medical imaging, where obtaining large annotated datasets for rare conditions can be extremely challenging. An intriguing aspect of SuPreM is its ability to learn a broader concept of ”objectness” through full supervision. This allows the model to transfer knowledge effectively to novel anatomical structures, even when these structures were not part of the pre-training set [15]. This generalization ability suggests that SuPreM is learning fundamental principles of anatomical segmentation, rather than just memorizing specific structures. The computational efficiency of SuPreM extends beyond pre-training to the fine-tuning phase as well. When fine-tuned on downstream tasks, SuPreM converges faster and achieves better performance than self-supervised models, reducing GPU hours needed from 60 to 20 in some experiments [15]. This efficiency makes SuPreM more accessible and practical for researchers and clinicians who may not have access to extensive computational resources.

2.4 Summary

Medical imaging is an indispensable tool in modern healthcare, offering a range of techniques to visualize the internal structures and functions of the body. Each imaging modality has its strengths and limitations, and the choice of technique depends on the specific clinical scenario. By providing detailed and accurate images, medical imaging facilitates early diagnosis, guides treatment decisions, and monitors the effectiveness of interventions, ultimately improving patient outcomes.

Rapid progress in medical imaging has coincided with significant advances in artificial intelligence, particularly in the field of deep learning. This convergence has given rise to a new paradigm in medical image analysis: foundation models. These models represent a significant leap forward from traditional task-specific algorithms, offering a more versatile and powerful approach to addressing a wide range of challenges in medical imaging.

Foundation models have shown immense potential in advancing medical image analysis by leveraging large-scale datasets and self-supervised learning techniques. These models, built upon architectures like Transformers and their variants, have demonstrated the ability to learn transferable representations that can be effectively adapted to various downstream tasks such as segmentation, classification, and anomaly detection.

The seminal works discussed in this chapter highlight the importance of carefully designed pretext tasks, model architectures, and large-scale heterogeneous pretraining datasets in developing powerful foundation models for medical imaging. These works have laid the groundwork for further research and development in this domain, showcasing the potential of self-supervised learning to leverage vast amounts of unannotated data.

However, challenges related to interpretability, explainability, computational resources, and effective adaptation to specific clinical settings or rare pathologies need to be addressed to ensure the responsible and reliable deployment of foundation models in medical imaging. As research in this area continues to advance, we can expect further innovations in adaptation strategies that enable the development of accurate, efficient, and interpretable medical imaging models based on foundation models.

In conclusion, foundation models represent a significant leap forward in medical image analysis, offering a versatile and powerful approach to addressing a wide range of challenges. With ongoing research and development, these models have the potential to revolutionize the field of medical imaging and ultimately improve patient outcome

Chapter 3

Methodology and Implementation

This chapter presents the methodological framework and experimental design employed in this study. It details the approaches used for data preparation, model development, and evaluation procedures, with specific attention to reproducibility and experimental validity in the context of medical image segmentation.

Model	**GitHub Repository**
MIS-FM	https://github.com/openmedlab/MIS-FM
DAE	https://github.com/Project-MONAI/research-contributions/tree/main/DAE
SuPreM	https://github.com/MrGiovanni/SuPreM
STU	https://github.com/uni-medical/STU-Net/tree/main
SPAD-Nets / PUMIT	https://github.com/function2-llx/PUMIT/tree/master
SwinMM	https://github.com/UCSC-VLAA/SwinMM/tree/master

Table 3.1: Model GitHub Repositories

3.1 Tasks specifications and Data Collection

In the pursuit of evaluating the capabilities of the foundation models in medical image analysis, it is imperative to select downstream tasks that not only possess clinical relevance but also rigorously test the postulated advantages of these models. To this end, we have identified two challenging segmentation tasks: AortaSeg24[2] and HaN-Seg[13]. These tasks were chosen for their complexity, clinical significance, and potential to elucidate the strengths and limitations of foundation models in medical imaging applications.

Fig. 3.1 and 3.2 provide visual representations of the desired segmentation masks and sample CT image slices for the AortaSeg24 and HaN-Seg challenges, respectively. These figures illustrate the intricate nature of the anatomical structures involved and the level of detail required for accurate segmentation.

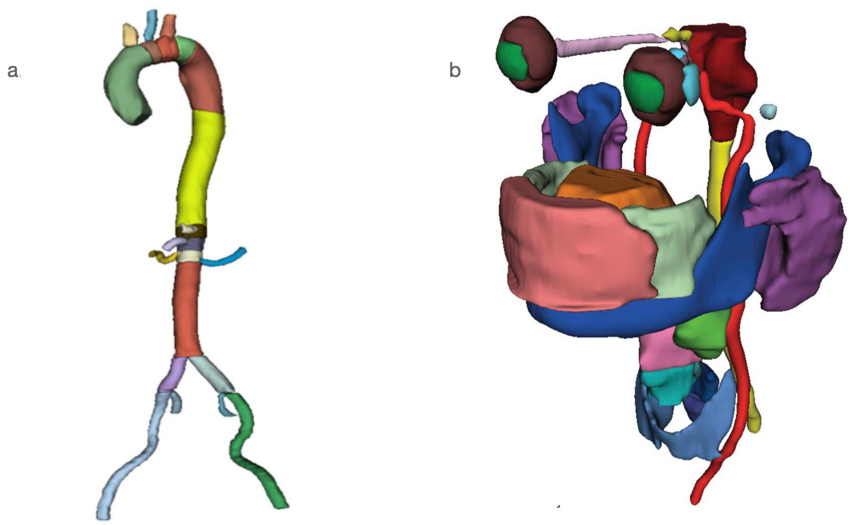

Figure 3.1: Ground Truth Segmentation maps of (a) AortaSeg24 and (b)HaN-Seg

3.1.1 AortaSeg24 Challenge

3.1.1.1 Task Description and Clinical Relevance

The AortaSeg24 challenge focuses on the segmentation of the aorta, its branches, and clinically relevant zones in Computed Tomography

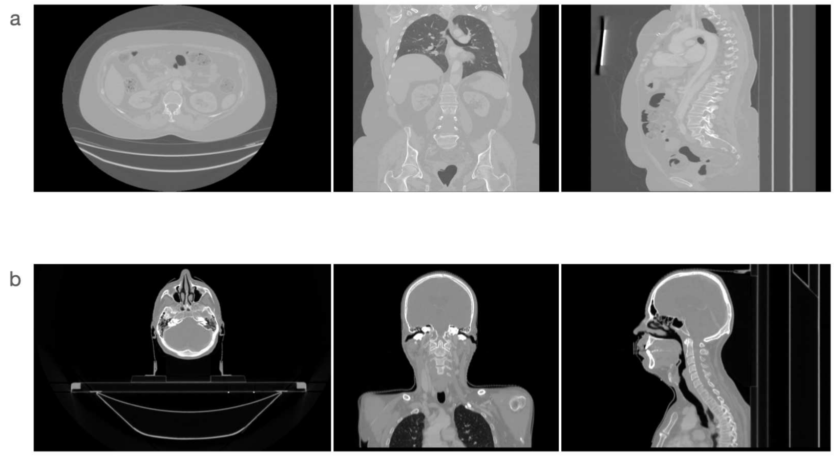

Figure 3.2: Slices of a sample of CT image from (a)AortaSeg24 and (b)HaN-Seg. From left-to-right the slices are along longitudinal, sagittal and frontal axis

Angiography (Computed Tomography Angiography (CTA)) images. This task necessitates the differentiation of multiple closely related vascular structures and a comprehensive understanding of their spatial relationships.

The aorta, which serves as the primary artery of the body, plays a pivotal role in the distribution of oxygenated blood throughout the body. Various aortic pathologies, including dissections, aneurysms, and atherosclerosis, present significant health risks and frequently require rapid medical intervention. Recent advances in imaging technologies, particularly CTA, have revolutionized the diagnosis and treatment of aortic diseases [2].

The shift towards endovascular procedures, which are now generally preferred over traditional surgical approaches, underscores the critical importance of precise anatomical information. Accurate multiclass segmentation of CTA images is indispensable for these minimally invasive techniques, which allows a detailed analysis of aortic structures and facilitates appropriate device selection for procedures involving the aorta and its branches [2].

3.1.1.2 Dataset Specification

The AortaSeg24 dataset comprises 50 CTA images designated for training purposes. These images exhibit considerable variability in their dimensions:

- Axial sizes range from 389x389 to 516x516 pixels
- The number of axial slices varies from 578 to 801 per image

Each CTA image is paired with meticulous annotations that delineate the aorta, its branches, and the clinically relevant SVS/STS zones. This comprehensive labeling enables the development and evaluation of advanced segmentation algorithms capable of handling the complex vascular anatomy.

3.1.1.3 Challenges and Limitations

The task of segmenting the aorta and its branches presents several non-trivial challenges:

- Differentiation of closely related vascular structures: The proximity and similarity of various blood vessels necessitate highly discriminative segmentation techniques.
- Accommodation of variations in image quality and contrast: CTA images may exhibit inconsistencies in contrast enhancement and overall quality, requiring robust algorithms capable of adapting to these variations.
- Accounting for anatomical variations among patients: The aorta and its branches can display significant inter-patient variability, demanding segmentation approaches that can generalize across diverse anatomical presentations.

These challenges collectively test the adaptability and precision of foundation models in handling complex, variable anatomical structures.

3.1.2 HaN-Seg challenge

3.1.2.1 Task Description and Clinical Relevance

The HaN-Seg challenge focuses on the segmentation of 30 organs-at-risk (OARs) in the head and neck region, utilizing both CT and MR imaging modalities. This anatomical area is characterized by its complexity, featuring numerous small, intricate structures in close proximity. Head and neck cancer represents a significant health concern, with radiotherapy being a primary treatment modality. This therapeutic approach necessitates precise targeting of cancerous cells while simultaneously protecting surrounding healthy tissues. Accurate three-dimensional segmentation of target volumes and OARs is crucial for optimal radiation dose planning [13].

Traditionally, computed tomography (CT) images have been the primary modality for radiotherapy planning. However, many OARs in the head and neck region exhibit superior visibility in magnetic resonance (MR) images. While previous research has explored Organs at Risk (OAR) segmentation from MR images in isolation, the potential synergistic benefits of combining CT and MR image analysis for this task remain underexplored. The Head and Neck Organ-at-Risk Multi-Modal Segmentation Challenge aims to catalyze the development and application of automated techniques that leverage multi-modal imaging data to enhance the accuracy of OAR segmentation in this critical anatomical region [13].

3.1.2.2 Dataset Specifications

The HaN-Seg dataset is comprised of:

- Paired CT and MR scans from 42 patients
- Corresponding ground truth 3D binary segmentation masks for OARs in the CT images

This multi-modal approach provides a rich dataset for developing and evaluating segmentation algorithms capable of integrating information from complementary imaging modalities.

3.1.2.3 Challenges and Limitations

The task of segmenting organs in the head and neck region using dual modalities presents several significant challenges:

- Integration of multi-modal information: Aligning and synthesizing data from both CT and MR modalities requires sophisticated techniques to effectively leverage the complementary information provided by each imaging method.
- Differentiation of structures with similar composition: The head and neck region contains numerous structures with similar tissue densities or intensities, necessitating advanced algorithms capable of discerning subtle differences.

- Adaptation to structural intricacies: The complex and variable anatomy of the head and neck region demands segmentation approaches that can adjust to the intricate and often patient-specific arrangements of OARs.

  These challenges collectively test the capability of foundation models to integrate multi-modal data and handle the complex, variable anatomical structures present in the head and neck region.

The AortaSeg24 and HaN-Seg challenges provide a comprehensive and rigorous framework for evaluating the performance of foundation models in medical image segmentation. These tasks not only assess the models' ability to handle complex anatomical structures but also their potential to significantly impact clinical workflows and improve patient care. By addressing these challenges, we aim to advance the state-of-the-art in medical image analysis and contribute to the development of more accurate and robust segmentation techniques for critical anatomical regions.

3.2 Model Implementation and Fine-tuning

3.2.1 Pre-processing

Across all models, we implement a core set of transformations from MONAI[39]. The process begins with image loading and channel arrangement, using `LoadImage` to load image data, followed by `EnsureChannelFirst` to standardize the channel dimension. This standardization is crucial for maintaining consistency across different datasets and model architectures. Spacing normalization is another critical step in our pipeline. We apply `Spacing` to resample images at a standard voxel size, typically (1.5, 1.5, 2.0) mm. This ensures consistent spatial relationships across datasets, which is particularly important for models that rely on learning spatial features.

Intensity normalization is achieved through two main approaches. For models like STU-Net[14] and SWIN-MM[9], we use `ScaleIntensity` to scale intensity values to a predefined range. This operation ensures that the intensity values are within a consistent range across all images, which can help faster convergence during training. For other models, such as MIS-FM[7] and our custom U-Net, we implemented a custom `AdvancedNormalization` transform. This performs foreground cropping, percentile-based clipping, and z-score normalization, which is particularly useful for focusing the model's attention on relevant parts of the image and handling outlier intensity values.

To ensure consistent image dimensions, we use spatial padding through `SpatialPad`. This is crucial for efficient batching during training and is often required by specific model architectures. Furthermore, we use random cropping via `RandCropByPosNegLabel` to extract balanced random patches during training. This technique is particularly important in medical image segmentation, where regions of interest often occupy only a small portion of the entire image volume.

Although these transformations form the core of our pipeline, each model incorporates additional specialized transforms. The STU-Net[14] and DAE[6] models employ a series of random augmentations that includes flips, rotations, and intensity adjustments. These enhance the models' robustness to common variations in medical images, such as orientation changes and intensity fluctuations.

The SuPreM model[15] introduces more advanced augmentation techniques. It applies random affine transformations, adds Gaussian noise, performs Gaussian smoothing, and adjusts image contrast. In particular, it includes a `SimulateLowResolution` transform, which prepares the model to handle inputs of varying quality, a common scenario in clinical settings.

The SPAD-Nets model[8] employs the most extensive augmentation pipeline in our study. In addition to the transforms used in SuPreM, it incorporates a `OneOf` transform that randomly applies either no transformation, gamma correction, or inverted gamma correction. This comprehensive approach aims to create a highly robust model by exposing it to a wide range of potential variations in the input data. The MIS-FM[7] and SWIN-MM[9] models use transform pipelines similar to STU-Net, but with some variations in spatial sizes and augmentation probabilities. These adjustments are made in accordance with the specific architectural requirements of these models. Our custom U-Net model demonstrates that effective results can be achieved with a relatively simple augmentation strategy. It uses a simplified pipeline focusing on basic augmentations such as flips and rotations, combined with our custom `AdvancedNormalization` technique.

The data transformation techniques described in this subsection are integral to the performance and generalization capabilities of models, and it is crucial to maintain consistency in the data transformation pipeline between the pre-training and fine-tuning stages. Using the same set of transforms ensures that the input data distribution remains consistent, allowing the model to leverage its pre-trained knowledge effectively during fine-tuning. This consistency helps preserve the learned features and representations from the pre-training phase, facilitating more efficient and stable fine-tuning on specific tasks.

3.2.2 Optimizers and Learning rate schedules

For our experiment, we employ a consistent optimization strategy across most models, with some variations for specific architectures. This approach ensures a fair comparison while allowing for model-specific optimizations.

3.2.2.1 Optimizer

We primarily use the AdamW optimizer[40], which combines the benefits of Adam's adaptive learning rates with weight decay regularization:

$$\theta_{t+1} = \theta_t - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_t \right) \tag{3.1}$$

where θ_t are the parameters at time t, η is the learning rate, $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moment estimates, ϵ is a small constant for numerical stability, and λ is the weight decay factor.

3.2.2.2 Learning Rate Schedulers

For most models, we implement a LinearWarmupCosineAnnealingLR schedule[41], which combines linear warmup with cosine annealing:

$$\mathrm{LR}(t) = \begin{cases} \mathrm{LR}_{\text{start}} + t \cdot \frac{\mathrm{LR}_{\text{base}} - \mathrm{LR}_{\text{start}}}{t_{\text{warmup}}} & \text{if } t < t_{\text{warmup}} \\ \mathrm{LR}_{\text{min}} + \frac{1}{2}(\mathrm{LR}_{\text{base}} - \mathrm{LR}_{\text{min}}) & \\ \quad \cdot (1 + \cos(\pi \cdot \frac{t - t_{\text{warmup}}}{t_{\text{max}} - t_{\text{warmup}}})) & \text{otherwise} \end{cases} \tag{3.2}$$

where t is the current epoch, LRstart is the initial learning rate for warmup, LRbase is the base learning rate, LRmin is the minimum learning rate, $t$warmup is the number of warmup epochs, and t_{max} is the total number of epochs.

3.2.2.3 Model-specific Variations

While most models use the AdamW optimizer and LinearWarmupCosineAnnealingLR scheduler, there are some exceptions as we adhere to instructions provided in respective repos:

STU-Net: Uses SGD optimizer with momentum and a polynomial learning rate decay[42]:

$$\mathrm{LR}(t) = \mathrm{LRinitial} \cdot (1 - \frac{t}{t\mathrm{max}})^{0.9} \tag{3.3}$$

SwinMM: Implements layer-wise learning rate decay using the `param_groups_layer_decay` function from the timm library[43].

3.3 Experimental Design and execution

Our experimental design is tailored to rigorously evaluate the generalization capabilities of foundation models on two downstream fine-grained tasks in medical image segmentation: AortaSeg24[2] and HaN-Seg[13]. The primary objective is to assess how well these pre-trained models can adapt to specific, detailed segmentation tasks with limited data, compared to a baseline model trained from scratch.

The objective of these experiments is to test how effectively foundation models generalize to fine-grained downstream tasks in medical image segmentation, particularly in few-shot learning scenarios?

Listed below are the six foundation models that we employ to conduct the experiments:

- STU-Net[14]
- SuPreM(SwinUNETR branch only)[15]
- DAE[6]
- SPAD-Nets[8]
- MIS-FM[7]
- SwinMM[9]

3.3.1 Evaluation Metrics

To comprehensively assess the performance and efficiency of the models, we employ a range of evaluation metrics:

- Dice Coefficient: Calculated for each class (0 to N, where 0 represents the background and N is the number of foreground classes) to measure segmentation accuracy[44].

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{3.4}$$

- Jaccard Index: Also known as Intersection over Union (IoU), computed for each class to provide a complementary measure of segmentation overlap[44].

$$JI = \frac{TP}{TP + FP + FN} \tag{3.5}$$

- Average Dice Coefficient: Computed across all classes to provide an overall measure of segmentation performance.
- DiceCE Loss: Recorded for both training and validation to monitor the learning process and assess overfitting.
- Samples per second: Measured to evaluate the processing speed of each model.

3.3.2 Baseline Comparison

To address why we would include a comparatively small Vanilla U-Net architecture in this study, the following interesting details are useful:

- Baseline Performance Benchmark: The small Vanilla U-Net acts as a fundamental baseline, providing a reference point against which the performance of more complex foundation models can be measured. This baseline is crucial for quantifying the advancements made by sophisticated architectures in medical image segmentation tasks.
- Computational Efficiency Comparison: By including a lightweight model alongside more parameter-heavy architectures, we can directly assess the trade-offs between model complexity and performance. This comparison is vital for understanding the efficiency gains offered by larger models relative to their increased computational demands.
- Real-world Clinical Applicability: We conducted in-house inference tests on a device that reflects the accessible compute resources available to individual clinicians. With the small Vanilla U-Net, the entire

process, including pre-processing and inference, takes approximately 3-4 minutes. This real-world performance data provides crucial insights into the practical applicability of different model sizes in clinical settings where computational resources may be limited and processing time is a critical factor.

- Resource-Constrained Scenarios: The performance of the small Vanilla U-Net on standard clinical hardware demonstrates its potential utility in settings where more powerful computational resources are not available. This addresses the practical considerations of model deployment across diverse healthcare environments, including smaller clinics or remote locations.

- Architecture Impact Analysis: The performance gap between the simple Vanilla U-Net and the more advanced models serves to highlight the significant impact of architectural innovations and pre-training strategies employed in foundation models, while also considering the trade-offs in terms of inference time and hardware requirements.

- Learning Dynamics Insight: Observing how quickly the Vanilla U-Net reaches its performance ceiling compared to larger models provides valuable insights into the learning dynamics and data efficiency of different architectures.

By incorporating this smaller model and providing concrete data on its real-world performance, we aim to offer a comprehensive and nuanced view of the current landscape in medical image segmentation. This approach spans from simple, lightweight architectures to complex, highly parameterized foundation models, while maintaining a focus on practical clinical applicability. It ensures a thorough analysis of the strengths and limitations of various model architectures in addressing the challenges posed by the AortaSeg24 and HaN-Seg tasks, taking into account both theoretical performance metrics and real-world deployment considerations.

Chapter 4

Results and Analysis

In this chapter, we present a comprehensive analysis of the performance of various foundation models and a baseline Vanilla U-Net in two challenging medical image segmentation tasks: Aortaseg24 and HaN-Seg. Our evaluation encompasses multiple dimensions, including segmentation accuracy, learning dynamics, class-wise performance, and computational efficiency.

4.1 Model Architecture and Computational Efficiency

To understand the practical implications of deploying foundation models in clinical settings, we first analyze their computational requirements and efficiency. Our experimental setup utilized a single NVIDIA A100 GPU with 40GB memory, running all experiments under consistent conditions to ensure fair comparison.

We measure computational efficiency through two key metrics:

1. Processing speed (samples processed per second during training)
2. Model size (number of parameters)

These metrics are particularly relevant for clinical deployment, where both real-time performance and hardware constraints play crucial roles. Table 4.1 summarizes the architectural characteristics and computational performance of each model.

The relationship between model size and processing speed reveals interesting patterns. SPAD-Nets, despite having the largest parameter count (134 million), achieves the second-highest processing speed (2.70 ± 0.60

Model	Patch Size(in voxels)	Parameters(in Millions)	Samples/sec
MIS-FM	64 × 128 × 128	45.30	1.40 ± 0.20
SPAD-Nets	96 × 96 × 96	134.00	2.70 ± 0.60
SwinMM	64 × 64 × 64	73.11	1.38 ± 0.23
DAE	96 × 96 × 96	62.89	2.99 ± 0.55
SuPreM	96 × 96 × 96	62.19	0.58 ± 0.16
Vanilla U-Net	128 × 128 × 128	0.47	0.13 ± 0.04
STU-Net	128 × 128 × 128	58.18	1.40 ± 0.23

Table 4.1: Architectural specifications and computational performance metrics of foundation models and baseline U-Net. Processing speed measured on NVIDIA A100 GPU, reported as mean(standard deviation) across atleast 1000 iters from Aortseg24 and Han-Seg training runs.

samples/second). This efficiency can be attributed to its spatially adaptive convolution operations that dynamically adjust computational resources based on input characteristics. DAE demonstrates the highest throughput (2.99 ± 0.55 samples/second) while maintaining a moderate parameter count (62.89 million), suggesting effective architectural design choices in balancing model capacity with computational efficiency.

The baseline Vanilla U-Net, while having the smallest parameter count (0.47 million), shows limited throughput (0.13 ± 0.04 samples/second). This lower processing speed stems from its basic architectural design that doesn't incorporate modern optimization techniques like adaptive computations or efficient attention mechanisms found in foundation models. The results highlight how architectural innovations in foundation models contribute to both model capacity and computational efficiency.

These findings have important implications for clinical deployment, where the choice between models must balance segmentation accuracy (discussed in subsequent sections) with computational constraints of medical facilities. The wide range of processing speeds and parameter counts suggests the need for careful consideration of available computational resources when selecting models for specific clinical applications.

4.2 Segmentation Performance and Learning Dynamics

Having established the computational characteristics of our models, we next investigate their learning capabilities and segmentation performance. This analysis bridges our understanding between computational costs and actual clinical utility, while setting up the foundation for more detailed class-wise analysis.

We examine learning behavior through two complementary lenses:

1. Efficiency of knowledge transfer from pre-training to medical tasks

2. Relationship between computational resources (discussed in Section 4.1) and segmentation quality

Our evaluation uses the Dice coefficient during training across two datasets with distinct computational and learning challenges:

- AortaSeg24: Multi-class vascular segmentation requiring precise boundary delineation

- HaN-Seg: Complex multi-organ segmentation with varying structure sizes and relationships

In the AortaSeg24 task (Figure 4.1(a.)), our experiments show interesting relationships between computational efficiency and segmentation performance. STU-Net, despite its moderate computational footprint (58.18M parameters, 1.40 samples/second), reaches a Dice score of 0.76. SwinMM follows with 0.68, suggesting that raw computational power doesn't directly translate to segmentation quality.

The HaN-Seg task (Figure 4.1(b.)) reveals how architectural choices influence performance on complex anatomical structures. The learning curves show greater volatility, particularly for models like MIS-FM and DA, which demonstrated high computational efficiency in Section 4.1. This suggests that efficient processing doesn't necessarily guarantee stable learning in complex anatomical scenarios.

The baseline Vanilla U-Net's performance (0.37 and 0.36 for AortaSeg24 and HaN-Seg respectively) provides context for both computational and learning aspects. Despite its minimal resource requirements (0.47M parameters), its learning trajectory suggests that architectural sophistication, rather than just parameter count, plays a crucial role in medical image segmentation.

Several patterns emerge that warrant deeper class-wise analysis:

1. Resource-Performance Balance: Higher computational efficiency doesn't consistently correlate with better segmentation performance

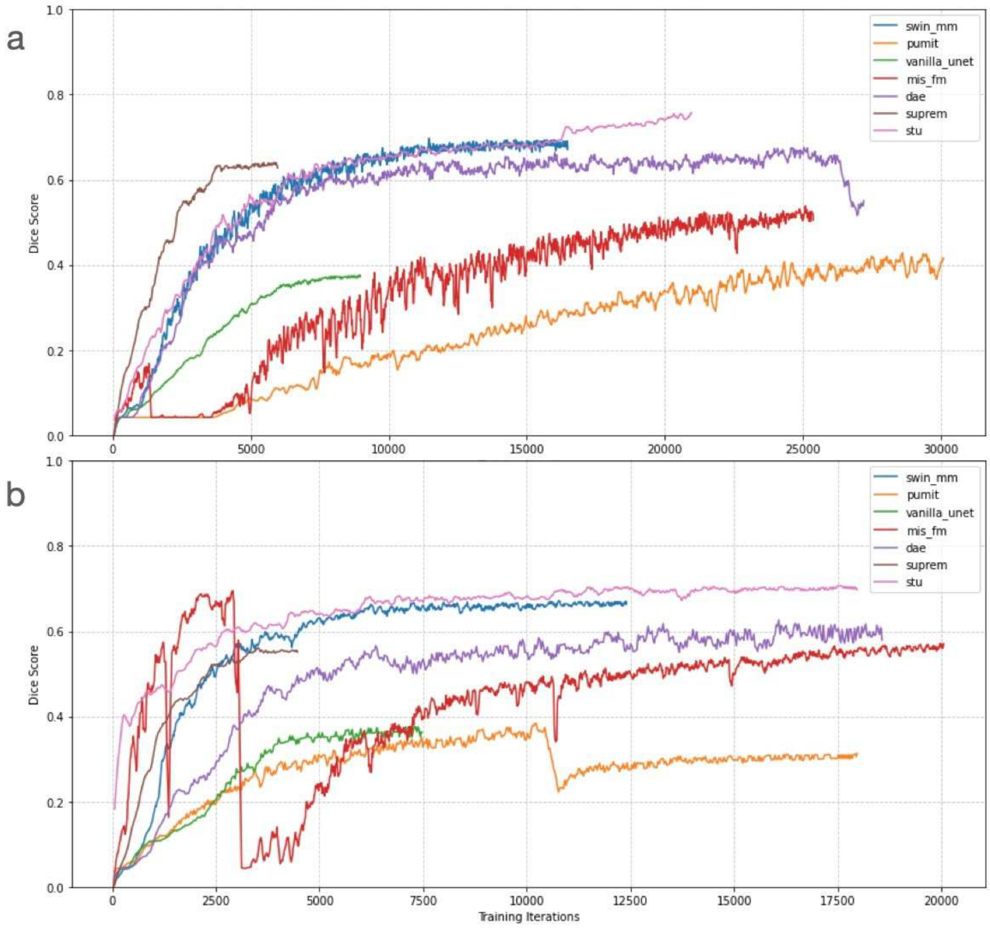

Figure 4.1: Learning curves showing Dice coefficient evolution during training for (a) AortaSeg24 and (b) HaN-Seg tasks. Curves represent single training runs for each model.

2. Task-Specific Adaptability: Models show varying abilities to adapt their pre-trained features to different anatomical contexts

3. Learning Stability: More complex architectures tend to show faster initial learning but with varying stability patterns

These observations raise important questions about performance across different anatomical structures, which we explore in detail through class-wise analysis in Section 4.3. The varying performance patterns also suggest that different architectural approaches might be optimal for different types of anatomical structures, a hypothesis we examine through detailed performance breakdown.

4.3 Performance Analysis Across Anatomical Scales

Following our examination of learning dynamics, we analyze how model performance varies across different anatomical scales. This analysis bridges our understanding of general learning behavior with specific segmentation challenges that become apparent in the qualitative analysis of 3D reconstructions presented in the next section.

We categorize anatomical structures based on their volume relative to the total segmentation mask:

- Minimal Volume Structures (V ≤ 1% Vtotal)
- Small Volume Structures (1% < V ≤ 5% Vtotal)
- Medium Volume Structures (5% < V ≤ 15% Vtotal)
- Large Volume Structures (V > 15% Vtotal)

Figure 4.2 presents the distribution of Jaccard scores across these volumetric groups for both AortaSeg24 (a) and HaN-Seg (b) tasks. The box plots show the median, quartiles, and outliers for each model within each volume category.

In the AortaSeg24 task, several interesting patterns emerge:

1. Volume-Related Patterns
    - Minimal volume structures (≤1%) show wide performance ranges (0.0-0.7)
    - Medium and large structures (>5%) exhibit more concentrated distributions, typically above 0.6
    - Outliers appear more frequently in smaller volume categories
    - The interquartile ranges generally narrow with increasing structure size
2. Model Behavior Patterns

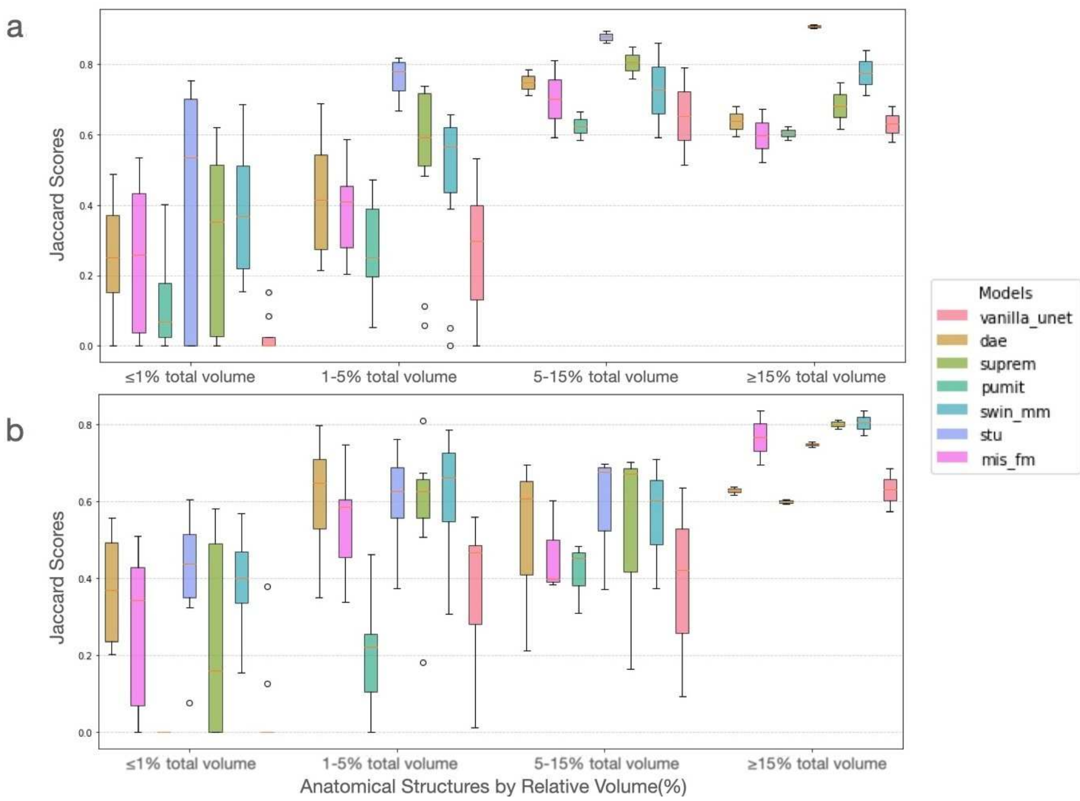

Figure 4.2: Distribution of Jaccard indices across anatomical scales for (a) AortaSeg24 and (b) HaN-Seg tasks. Box plots show median (horizontal line), interquartile range (box), and outliers (points) for each model across volume-based structure groups.

- STU-Net's distributions show relatively consistent median values across categories
- The baseline Vanilla U-Net exhibits wider interquartile ranges in most categories
- SPAD-Nets and MIS-FM display broader performance distributions for minimal volume structures
- Models tend to show reduced variability in their performance as structure size increases

The HaN-Seg task presents different distribution patterns:

1. Task-Specific Distributions

- Broader interquartile ranges across all volume categories
- More frequent outliers in medium-sized structures
- Less pronounced relationship between structure size and score distributions
- Greater overlap in performance distributions between adjacent volume categories

2. Model Distribution Characteristics

- MIS-FM shows progressively narrowing distributions with increasing structure size
- SwinMM exhibits compact distributions for larger structures but wide ranges for minimal volumes
- Distribution shapes vary more between models compared to AortaSeg24
- Outlier behavior appears more pronounced in smaller volume categories

These observed patterns provide context for understanding the 3D reconstruction results presented in Section 4.4. The distribution characteristics across volume categories help explain specific features we observe in the reconstructed anatomical structures, particularly for minimal and small volume structures where performance distributions show greater variability.

The relationship between structure size and score distributions appears across both tasks, though with task-specific characteristics. This observation informs our understanding of the qualitative differences in 3D reconstructions between AortaSeg24 and HaN-Seg tasks, which we examine in detail in the following section.

4.4 Computational Requirements and Training Dynamics

Understanding the practical implications of deploying foundation models in clinical settings requires analysis of their computational demands. Our experimental setup utilized a single NVIDIA A100 GPU with 40GB memory, running all experiments under consistent conditions to ensure comparable measurements.

We examine computational requirements through training iterations needed for convergence - a key indicator of training efficiency and resource utilization in practical deployment scenarios. Figure 4.3 illustrates the number of training iterations required for each model across both AortaSeg24 and HaN-Seg tasks.

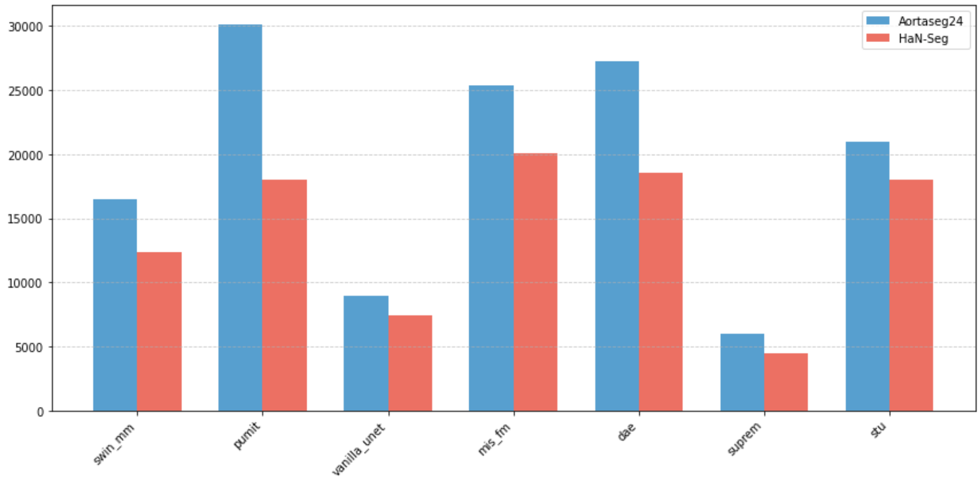

Figure 4.3: Training iterations required for convergence across foundation models and baseline U-Net. Blue bars represent AortaSeg24 task, red bars represent HaN-Seg task. Numbers above bars indicate exact iteration counts.

Several interesting patterns emerge from this analysis:

1. Task-Specific Training Requirements

    - Models consistently require more iterations for AortaSeg24 than HaN-Seg
    - The difference in required iterations varies substantially across models
    - SPAD-Nets shows the largest disparity (30,090 vs 17,970 iterations)
    - SuPreM demonstrates the smallest difference (5,970 vs 4,470 iterations)
    - The complexity of vascular structures in AortaSeg24 might contribute to increased iteration requirements

2. Model-Specific Characteristics

    - SuPreM achieves convergence with notably fewer iterations (5,970/4,470) across both tasks
    - SPAD-Nets requires the highest number of iterations (30,090/17,970)
    - The baseline Vanilla U-Net shows moderate iteration requirements (8,970/7,470)
    - Most foundation models need 15,000-30,000 iterations for AortaSeg24
    - Models with similar architectures (like STU-Net and MIS-FM) show comparable iteration requirements

3. Cross-Task Patterns

    - Models maintain relatively consistent ranking of iteration requirements across tasks
    - The ratio between AortaSeg24 and HaN-Seg iterations remains roughly similar for most models
    - Foundation models generally require 1.5-2x more iterations for AortaSeg24 compared to HaN-Seg
    - The consistent ratio suggests inherent task complexity differences rather than model-specific issues

4. Efficiency Considerations

    - No clear correlation appears between model size and required iterations
    - Pre-training strategies seem to influence convergence speed more than architecture size
    - SuPreM's efficient convergence might reflect its supervised pre-training approach
    - Models with self-supervised pre-training (like SPAD-Nets) typically require more iterations
    - The baseline Vanilla U-Net's moderate iteration count suggests architecture complexity doesn't necessarily predict training efficiency

These observations reveal varying levels of training efficiency across models and tasks. SuPreM's rapid convergence might indicate effective pre-training for medical image segmentation tasks, while SPAD-Nets's higher iteration requirements could reflect its architectural complexity. The consistent pattern of higher iterations for AortaSeg24 suggests task-specific challenges in vascular segmentation.

The relationship between training iterations and segmentation performance becomes particularly interesting when examining learning dynamics. As we explore in the next section, models requiring fewer iterations, like SuPreM, don't necessarily show faster learning rates or better final performance. This disconnect between computational efficiency and segmentation accuracy raises important questions about the trade-offs between training resources and model performance, which we investigate through detailed learning curves and convergence patterns.

4.5 Learning Behaviour Across Models and Tasks

After analyzing model architectures and computational requirements, we examine the learning dynamics revealed through training trajectories. Figure 4.4 presents paired learning curves for each model, comparing their performance on both AortaSeg24 and HaN-Seg tasks against our baseline Vanilla U-Net. The six subplots, arranged to facilitate direct comparison, reveal striking patterns in how foundation models leverage their pre-trained knowledge for medical image segmentation.

In AortaSeg24, examining the top-right subplot in Figure 4.4, STU-Net demonstrates remarkably efficient knowledge transfer. Its trajectory shows a steep initial ascent during the first 5,000 iterations, indicating effective utilization of pre-trained features for vascular structure recognition. In contrast, as seen in the bottom-middle subplot, SuPreM exhibits a different but equally efficient learning pattern - while its ascent is more gradual, it achieves stable performance in fewer total iterations, typically converging around 6,000 iterations compared to other models' 15,000-20,000 iterations.

The dotted lines in each subplot, representing the baseline Vanilla U-Net's behavior, provide a crucial reference point. The baseline's consistently lower trajectory and earlier performance plateau (around 0.37 Dice score) demonstrate the limitations of learning from scratch. This performance gap, clearly visible across all subplots, remains consistent across both tasks, with

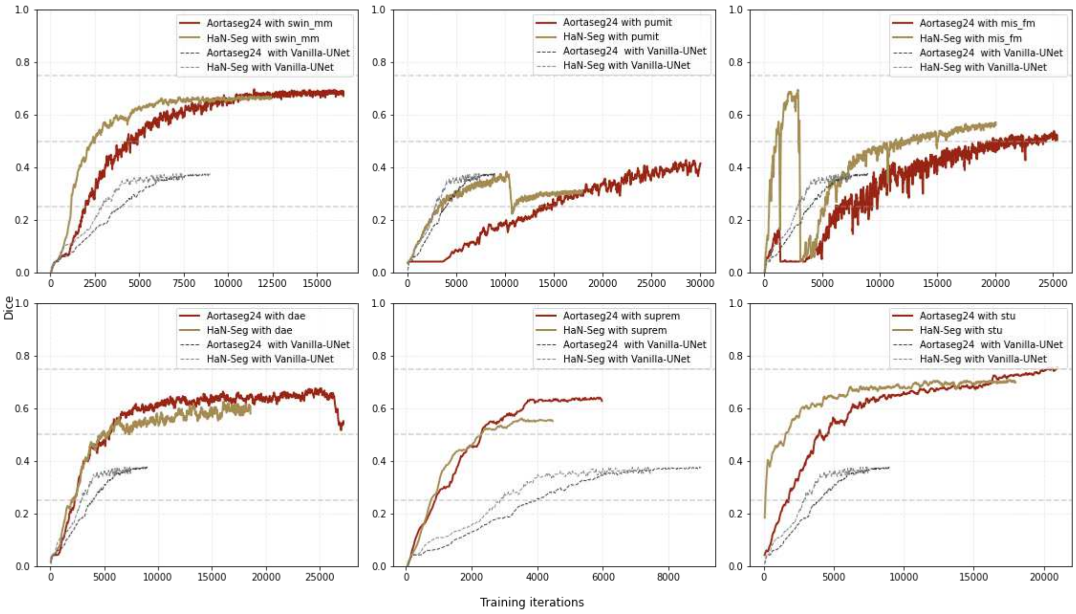

Figure 4.4: Comparative learning trajectories showing training dynamics of foundation models and baseline Vanilla U-Net across tasks. Each subplot presents paired learning curves for a specific model (red: AortaSeg24, brown: HaN-Seg) against the baseline (dotted lines), tracking Dice coefficient against training iterations. Top row (left to right): SwinMM shows consistent convergence patterns; SPAD-Nets demonstrates extended training requirements; MIS-FM exhibits notable performance fluctuations. Bottom row: DA maintains stability until late-stage training; SuPreM achieves rapid convergence with fewer iterations; STU-Net displays efficient knowledge transfer with steep initial learning. All foundation models consistently outperform the Vanilla U-Net baseline, though with varying stability and convergence characteristics. Curves represent single training runs, highlighting real learning dynamics rather than averaged behavior. Note the different x-axis scales, reflecting varying convergence times across models (ranging from 8,000 to 30,000 iterations). The consistent gap between foundation models and baseline (approximately 0.3-0.4 Dice score) demonstrates the advantage of pre-trained features in medical image segmentation tasks.

foundation models maintaining a 0.3-0.4 Dice score advantage throughout training.

Stability characteristics vary significantly among models. The top-right subplot reveals MIS-FM's intriguing behavior with noticeable performance fluctuations, particularly evident in the HaN-Seg task where its brown curve shows a dramatic performance spike around iteration 5,000. These fluctuations suggest complex optimization dynamics, possibly related to the model's multi-scale feature learning approach. In contrast, as shown in the bottom-left subplot, DA maintains relatively stable learning curves until a sudden performance drop in later iterations of AortaSeg24 training, highlighting potential optimization challenges in maintaining learned features.

The HaN-Seg trajectories (brown curves across all subplots) reveal different aspects of model behavior. Looking at the top-left subplot, SwinMM demonstrates particularly strong performance on this task, suggesting its architecture may be better suited to handling multiple anatomical structures simultaneously. The more pronounced fluctuations during middle training phases (iterations 5,000-15,000), visible across multiple subplots, indicate ongoing adaptation to complex spatial relationships between different anatomical structures.

Task-specific adaptation patterns become evident when comparing the red (AortaSeg24) and brown (HaN-Seg) curves within each subplot. While models generally require more iterations to achieve stable performance on AortaSeg24, they show faster initial improvement on HaN-Seg. This difference, particularly clear in the STU-Net and SwinMM subplots, suggests that features learned during pre-training may be more directly applicable to general anatomical structures than to specialized vascular patterns.

Performance plateaus manifest differently across models and tasks, as evidenced by the curve endpoints in Figure 4.4. In AortaSeg24 (red curves), most models show gradual improvement until late in training, while HaN-Seg trajectories (brown curves) tend to stabilize earlier. The bottom-middle subplot shows SuPreM's rapid convergence on both tasks, suggesting particularly efficient feature adaptation, while SPAD-Nets's extended training requirement in the top-middle subplot indicates a more complex optimization landscape.

The stability differences observed across subplots directly influence segmentation quality. Models exhibiting smoother learning curves, like STU-Net (top-right) and SuPreM (bottom-middle), tend to produce more consistent segmentation results. The performance fluctuations seen in models like MIS-FM correlate with variations in segmentation boundary consistency, particularly for smaller anatomical structures.

These detailed learning trajectories provide crucial insights into how foundation models adapt to medical segmentation tasks. The consistent superiority over the baseline, varying convergence patterns, and task-specific adaptations revealed in Figure 4.4 help explain the quantitative and qualitative differences observed in segmentation performance. Understanding these learning behaviors is essential for optimizing model selection and training strategies for specific clinical applications.

4.6 Qualitative Analysis of 3D Reconstructions

The qualitative analysis of 3D reconstructions provides crucial insights into the real-world performance of these models. Figure 4.5 presents the 3D reconstructions from each model for both Aortaseg24 (a) and HaN-Seg (b) tasks on test samples.

In the Aortaseg24 task (Figure 4.5a), STU-Net, SuPreM, and SwinMM demonstrate superior reconstruction quality, maintaining smooth vessel boundaries and accurate branching structures. These models successfully capture the complex topology of the aorta and its branches, with STU-Net particularly excelling in preserving fine details of smaller vessels. In contrast, MIS-FM shows significant fragmentation in its reconstruction, especially in the smaller branches, while SPAD-Nets's reconstruction lacks detail and appears oversimplified.

The HaN-Seg reconstructions (Figure 4.5b) reveal the increased complexity of head and neck anatomy. STU-Net and SwinMM achieve the most coherent reconstructions, accurately representing the spatial relationships between different organs-at-risk. DAE and MIS-FM maintain reasonable structural integrity but show some inconsistencies in organ boundaries. SPAD-Nets's performance is notably poor, failing to capture many anatomical structures and showing significant segmentation artifacts.

Interestingly, when these models were tested on MRI data, their performance degraded significantly. This degradation can be attributed to the fundamental differences in image characteristics between CT and MRI, and the inadequacy of CT-optimized pre-processing pipelines for MRI data. While CT images have standardized Hounsfield units and consistent intensity distributions, MRI intensities are arbitrary and can vary significantly across scanners and protocols. Our pre-processing pipeline, which includes intensity normalization and standardization techniques optimized for CT, fails to appropriately handle the unique characteristics of MRI data, resulting in poor segmentation performance.

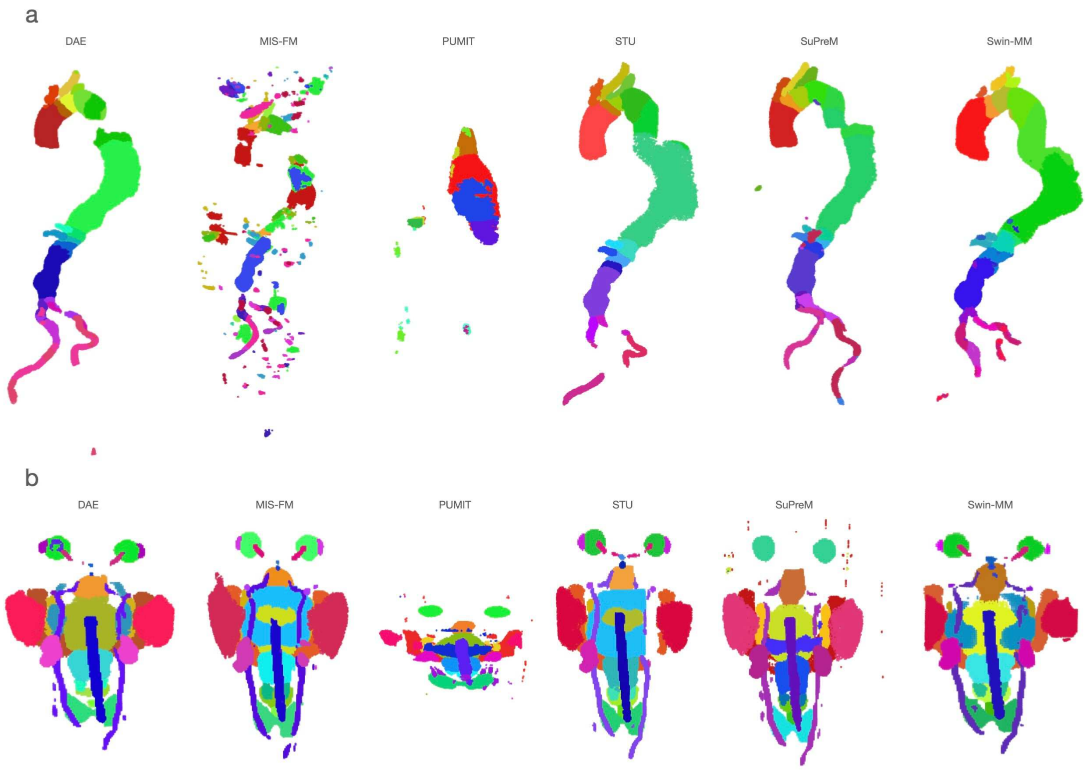

Figure 4.5: 3D reconstructions from different models for (a.) Aortaseg24 and (b.) HaN-Seg test samples

These observations highlight the importance of modality-specific pre-processing and potentially the need for specialized foundation models for different imaging modalities. The superior performance on CT data, particularly by models like STU-Net and SwinMM, suggests that these architectures have effectively learned the spatial and intensity patterns specific to CT imaging, but this specialization may come at the cost of cross-modality generalization.

Chapter 5

Discussion

Our comprehensive evaluation of foundation models against a baseline Vanilla U-Net on the Aortaseg24 and HaN-Seg tasks revealed several significant patterns. STU-Net consistently outperformed other models within the scope of our experiment, achieving Dice scores of 0.76 and 0.70 on Aortaseg24 and HaN-Seg respectively - a substantial improvement over the baseline's 0.37 and 0.36. This performance advantage was not wuniform across anatomical structures; the models excelled at segmenting larger, more distinct structures (with STU-Net achieving Dice scores above 0.90 for several classes in Aortaseg24) but struggled with smaller, intricate anatomical regions. Computational efficiency varied markedly among models. DAE processed 2.99 ± 0.55 samples per second, demonstrating high throughput, while SuPreM showed remarkable convergence speed, requiring only 6000 iterations for Aortaseg24. Conversely, SPAD-Nets needed 30120 iterations, highlighting significant variations in learning efficiency. These differences remained consistent across both tasks, suggesting inherent architectural characteristics rather than task-specific behaviors. An unexpected finding emerged in our cross-modality testing: models pre-trained on CT data showed significant performance degradation when applied to MRI data. This degradation persisted despite attempting various pre-processing adjustments, pointing to fundamental challenges in cross-modality generalization. Additionally, our qualitative analysis through 3D reconstructions revealed that quantitative metrics don't tell the complete story - models with similar Dice scores sometimes produced notably different reconstruction qualities.

5.1 Critical Reflections and Limitation

Several important limitations affect the conclusiveness of our study's outcomes. The most significant is the absence of statistical validation of our results. Without conducting multiple training runs with different random seeds, we cannot definitively assert the statistical significance of the performance differences observed between models. This limitation is particularly relevant given the inherent variability in deep learning model training. Our comparison using a single train-validation split, while computationally efficient, doesn't account for potential data distribution biases. Cross-validation would have provided more robust performance estimates and confidence intervals for our metrics. For instance, when STU-Net achieved a Dice score of 0.76 on Aortaseg24, we cannot quantify the stability of this performance across different data splits. The evaluation on clinical hardware was limited to high-end GPUs (NVIDIA A100). While we discuss implications for resource-constrained environments, we didn't actually test model performance on more modest hardware configurations typical in clinical settings. This gap between our testing environment and real-world deployment scenarios limits our ability to make concrete recommendations about practical applicability. Our preprocessing pipeline was optimized primarily for CT data, as evidenced by the poor performance on MRI data. The lack of a modality-agnostic preprocessing approach limits the generalizability of our findings. Additionally, we didn't systematically evaluate the impact of different preprocessing choices on model performance. The batch sizes and learning rates were chosen based on hardware constraints and common practices rather than through systematic hyperparameter optimization. This means we might not have observed the true optimal performance of each model. The computational constraints also limited our ability to test larger model variants or ensemble approaches that could potentially yield better results.

5.2 Scientific Contribution

Our work contributes several important findings to the ongoing discourse about foundation models in medical imaging. First, our systematic comparison of model architectures on identical tasks reveals that the relationship between model complexity and performance isn't straightforward. While STU-Net's superior performance (0.76 Dice score on Aortaseg24) aligns with previous findings about the benefits of large-scale pre-training[14], SPAD-Nets' higher parameter count (134M vs STU-Net's 58.18M) didn't translate to better performance, challenging assumptions about model scaling in medical imaging tasks.

The class-wise performance analysis advances our understanding of how these models handle anatomical structures of varying sizes. Previous works often reported aggregate metrics[9, 7], but our detailed analysis shows that even top-performing models struggle with smaller structures, suggesting a limitation in current architectural designs. This finding connects directly to ongoing discussions about attention mechanisms and multi-scale feature learning in medical imaging[8].

Our observation about computational efficiency variations - particularly SuPreM's rapid convergence (6000 iterations) compared to SPAD-Nets (30120 iterations) - contributes to the growing body of work on efficient training strategies[15]. This finding suggests that sophisticated pre-training strategies might be more crucial for efficiency than architectural complexity, a perspective that adds nuance to current debates about model design.

The degradation in performance when testing these CT-trained models on MRI data highlights a crucial gap in current approaches to medical image analysis. While previous works have explored cross-modality learning[6, 9], our results suggest that current solutions for handling modality-specific characteristics are insufficient, pointing to a critical area needing attention in the field.

These contributions collectively suggest that the field might benefit from shifting focus from raw model size to more nuanced architectural considerations that account for structure-specific challenges and practical deployment constraints. Our findings particularly resonate with recent work questioning the "bigger is better" paradigm in medical imaging AI[14, 15], while offering specific insights about where and why this might be the case.

5.3 Broader Implications

Our findings have several implications that extend beyond technical performance metrics. From an industrial perspective, the efficiency variations we observed among models (from DAE's 2.99 samples/second to Vanilla U-Net's 0.13 samples/second) directly impact deployment costs and scalability. Healthcare providers must balance segmentation accuracy against computational requirements, particularly in resource-constrained settings. The rapid convergence demonstrated by models like SuPreM suggests potential for reducing training time and associated energy costs, though this needs validation in real clinical environments.

The societal impact extends to healthcare accessibility. More efficient models could enable smaller clinics and hospitals to benefit from automated segmentation capabilities. However, the current performance-efficiency trade-off might exacerbate healthcare disparities if only well-resourced institutions can deploy the best-performing models. The varying computational requirements we observed (from SuPreM's efficient convergence to SPAD-Nets' longer training needs) suggest that thoughtful model selection based on available resources is crucial for equitable deployment.

5.3.1 Sustainability Considerations

From a sustainability perspective, the computational demands of these models warrant attention. Training foundation models requires significant computational resources - SPAD-Nets' 30120 iterations for convergence translates to substantial energy consumption. While smaller models like our Vanilla U-Net (0.47M parameters) require less compute, their lower performance creates a challenging trade-off between environmental impact and clinical utility. This balance becomes particularly relevant when considering widespread deployment across healthcare systems.

5.3.2 Ethical Concerns

Several ethical considerations emerge from our work. First, the inconsistent performance across different anatomical structures raises concerns about reliability - while STU-Net achieved high accuracy (>0.90 Dice score) for larger structures, the reduced performance on smaller structures could affect clinical decisions if not properly understood and communicated. This variability necessitates careful consideration of how these models are integrated into clinical workflows. Additionally, our models' poor generalization to MRI data highlights the risk of modality-specific biases that could affect patient care if not explicitly addressed.

These considerations collectively highlight the need for a holistic approach to model development that considers not just technical performance, but also practical constraints, ethical implications, and environmental impact. Future developments in this field should explicitly address these broader concerns alongside technical improvements.

Chapter 6

Conclusions and future work

This study set out to investigate the efficacy of foundation models in medical image segmentation, focusing specifically on CT scans for the Aortaseg24 and HaN-Seg tasks. Our comprehensive analysis has yielded several significant insights into the performance, adaptability, and computational efficiency of these models. Reflecting on our initial objectives, we have successfully implemented and fine-tuned a range of foundation models for medical image segmentation tasks, evaluating them on two challenging datasets: AortaSeg24 and HaN-Seg. Our results demonstrate that foundation models, particularly STU-Net and SwinMM, consistently outperform the baseline Vanilla U-Net across both tasks. This superior performance underscores the potential of transfer learning and sophisticated architectural designs in advancing medical image analysis. The positive effects of using foundation models are numerous. We observed rapid initial learning in models like STU-Net and SuPreM, indicating efficient knowledge transfer from pre-training. This quick adaptation could be invaluable in clinical scenarios where rapid model deployment is crucial. Furthermore, the ability of these models to handle complex, multi-structure segmentation tasks, as evidenced by their performance on HaN-Seg, highlights their potential to revolutionize medical imaging across various anatomical regions. However, our study also revealed several drawbacks and challenges. The task-specific variability in performance, particularly evident in models like PUMIT and MIS-FM, underscores the complexity of developing truly versatile foundation models for medical imaging. Additionally, the computational demands of these larger models, as compared to the lightweight Vanilla U-Net, raise important considerations about computational requirements that could affect their deployment in different clinical environments.

Our evaluation of results reveals a nuanced landscape. While our single-run experiments suggest potential advantages of foundation models over traditional architectures, the lack of multiple training runs and statistical validation means we cannot make definitive claims about performance superiority. The observed relationships between model size, computational efficiency, and segmentation accuracy provide preliminary insights that warrant further rigorous investigation through multiple training runs and proper statistical analysis. These initial observations serve as hypotheses for future studies rather than conclusive findings. The class-wise performance analysis, utilizing both Dice and Jaccard indices, provided valuable insights into the models' capabilities across different anatomical structures. This dual-metric approach highlighted the importance of using multiple evaluation criteria for a comprehensive assessment of segmentation quality. Looking towards the future, several avenues for further research emerge:

6.1 Future Work

While our experiments demonstrated the effectiveness of foundation models in segmentation tasks, their computational demands remain a practical concern. Future work should focus on model compression techniques and hardware-aware optimization strategies. Specifically, investigating approaches like knowledge distillation and quantization could help bridge the gap between model performance and deployment feasibility. Development of lightweight architectures that maintain segmentation accuracy while reducing memory footprint would be particularly valuable.

Our class-wise analysis revealed varying performance across different anatomical structures. Future research could explore architectural modifications that better handle both large, distinct structures and small, intricate ones within the same model. This might involve developing adaptive attention mechanisms or structure-specific processing pathways that can be dynamically engaged based on the target anatomy.

The significant performance degradation observed when applying CT-trained models to MRI data highlights a crucial area for improvement. Future work should investigate:

- Development of modality-agnostic preprocessing pipelines
- Architecture modifications that better handle intensity distribution variations

- Self-supervised learning approaches that can leverage unpaired multi-modal data

While our study focused on performance metrics and computational efficiency, the ”black box” nature of these models remains a concern for clinical adoption. Future research directions could include:

- Development of attention visualization techniques specific to 3D medical data
- Integration of uncertainty quantification methods
- Creation of interpretable decision paths for segmentation choices

Given the perpetual challenge of limited labeled medical data and model robustness, future work may also explore:

- Few-shot learning techniques optimized for volumetric data
- Semi-supervised approaches that effectively leverage unlabeled volumes
- Active learning strategies for efficient annotation selection
- Systematic evaluation of model performance across different imaging protocols
- Development of quality assessment metrics for segmentation reliability
- Methods for handling edge cases and anatomical variations

These research directions aim to address both technical limitations and practical deployment challenges identified in our study. Progress in these areas could contribute significantly in the clinical utility of foundation models in medical image analysis.

References

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021. [Pages 1 and 12.]

[2] M. Imran, J. R. Krebs *et al.*, "Multi-class segmentation of aortic branches and zones in computed tomography angiography," https://aortaseg24.grand-challenge.org/, 2024, accessed on September 16, 2024. [Pages 1, 3, 28, 29, and 35.]

[3] O. Ronneberger, P. Fischer, and A. Becker, "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015. [Page 2.]

[4] J. Zhang, X. Wang, and Z. Zhang, "Data augmentation techniques for deep learning in medical imaging: A review," *Journal of Medical Imaging*, vol. 9, no. 1, pp. 1–16, 2022. [Page 2.]

[5] W. Chen, L. Zhang, and J. Liu, "Transfer learning in medical image analysis: A survey," *IEEE Transactions on Medical Imaging*, vol. 42, no. 2, pp. 1234–1250, 2023. [Page 2.]

[6] J. M. J. Valanarasu, Y. Tang, D. Yang, Z. Xu, and C. Xie, "Disruptive autoencoders: Leveraging low-level features for 3d medical image pre-training," *arXiv preprint arXiv:2307.16896*, 2023. [Pages 2, 6, 14, 17, 18, 33, 35, and 54.]

[7] G. Wang, J. Wu, X. Luo, X. Liu, K. Li, and S. Zhang, "Mis-fm: 3d medical image segmentation using foundation models pretrained on a large-scale unannotated dataset," *arXiv preprint arXiv:2306.16925*, 2023. [Pages 3, 6, 17, 20, 21, 32, 33, 35, and 54.]

[8] L. Luo, X. Chen, B. Tang, X. Chen, R. Han, C. Hu, Y. Li, and T. Chen, "Building universal foundation models for medical image analysis with spatially adaptive networks," 2024. [Online]. Available: https://arxiv.org/abs/2312.07630 [Pages 3, 6, 18, 19, 20, 33, 35, and 54.]

[9] Y. Wang, Z. Li, J. Mei, Z. Wei, L. Liu, C. Wang, S. Sang, A. L. Yuille, C. Xie, and Y. Zhou, "Swinmm: masked multi-view with swin transformers for 3d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 486–496. [Pages 3, 6, 21, 22, 23, 32, 33, 35, and 54.]

[10] Y. Wang, F. Zhang, and X. Liu, "Versatile architectures for medical image segmentation: A comprehensive review," *Medical Image Analysis*, vol. 78, pp. 101–115, 2023. [Page 3.]

[11] W. Liu, Y. Wang, and J. Zhang, "Lightweight architectures for medical image segmentation: A comparative study," *Journal of Healthcare Engineering*, vol. 2022, 2022. [Page 3.]

[12] A. Kumar, R. Singh, and A. Gupta, "Adaptive loss functions for imbalanced data in medical imaging: A review," *Artificial Intelligence in Medicine*, vol. 130, 2023. [Page 3.]

[13] "Head and neck organ-at-risk segmentation challenge," https://han-seg2023.grand-challenge.org/han-seg2023/, 2023, accessed on September 16, 2024. [Pages 3, 28, 31, and 35.]

[14] Z. Huang, H. Wang, Z. Deng, J. Ye, Y. Su, H. Sun, J. He, Y. Gu, L. Gu, S. Zhang *et al.*, "Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training," *arXiv preprint arXiv:2304.06716*, 2023. [Pages 6, 16, 17, 32, 33, 35, and 54.]

[15] W. Li, A. Yuille, and Z. Zhou, "How well do supervised 3d models transfer to medical imaging tasks?" in *The Twelfth International Conference on Learning Representations*. [Pages 6, 23, 24, 33, 35, and 54.]

[16] S. Hussain, I. Mubeen, N. Ullah, S. S. U. D. Shah, B. A. Khan, M. Zahoor, R. Ullah, F. A. Khan, and M. A. Sultan, "Modern diagnostic imaging technique applications and risk factors in the medical field: a

review," *BioMed research international*, vol. 2022, no. 1, p. 5164970, 2022. [Page 8.]

[17] M. A. Haidekker, "Medical imaging technology," 2013. [Pages 9, 10, 11, and 12.]

[18] L. Antiga, "Patient-specific modeling of geometry and blood flow in large arteries," *Politecnico di Milano*, 2002. [Pages 9, 10, and 11.]

[19] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, S. McCandlish, G. Brockman, and O. Team, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165 [Page 12.]

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. [Page 12.]

[21] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020. [Page 12.]

[22] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019. [Page 13.]

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021. [Page 13.]

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. [Page 13.]

[25] C. Matsoukas, J. F. Haslum, M. Söderberg, K. Smith, M. Erdt, A. Jöud, and V. S. Pankratz, "Medical transformer: Universal brain encoder for 3d mri analysis," *arXiv preprint arXiv:2104.13633*, 2021. [Pages 13, 14, and 15.]

[26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021. [Page 13.]

[27] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar, "Moco pretraining improves representation and transferability of chest x-ray models," *arXiv preprint arXiv:2101.05224*, 2021. [Page 14.]

[28] Y. Zhou, Y. Zhang, X. Bai, X. Chen, X. Hou, A. Yuille, and P. Luo, "Masked autoencoders are scalable vision learners," *arXiv preprint arXiv:2111.06377*, 2022. [Page 14.]

[29] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen *et al.*, "Big self-supervised models advance medical image classification," *arXiv preprint arXiv:2101.05224*, 2021. [Pages 14 and 15.]

[30] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications*, vol. 11, no. 1, p. 3673, 2020. [Page 14.]

[31] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable machine learning in healthcare," *arXiv preprint arXiv:2011.08737*, 2020. [Page 14.]

[32] X. Hu, X. Xu, and Y. Shi, "How to efficiently adapt large segmentation model (sam) to medical images," *arXiv preprint arXiv:2306.13825*, 2023. [Page 15.]

[33] G. Deng, K. Zou, K. Ren, M. Wang, X. Yuan, S. Ying, and H. Fu, "Domain adaptation for medical image analysis: A survey," *arXiv preprint arXiv:2107.09011*, 2021. [Page 16.]

[34] D. M. H. Nguyen, H. Nguyen, N. T. Diep, T. N. Pham, T. Cao, B. T. Nguyen, P. Swoboda, N. Ho, S. Albarqouni, P. Xie, D. Sonntag, and M. Niepert, "Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching," *arXiv preprint arXiv:2305.19226*, 2023. [Page 16.]

[35] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021. [Page 16.]

[36] T. A. D'Antonoli, L. K. Berger, A. K. Indrakanti, N. Vishwanathan, J. Weiß, M. Jung, Z. Berkarda, A. Rau, M. Reisert, T. Küstner *et al.*, "Totalsegmentator mri: Sequence-independent segmentation of 59 anatomical structures in mr images," *arXiv preprint arXiv:2405.19492*, 2024. [Page 16.]

[37] Y. Jiang, M. Sun, H. Guo, X. Bai, K. Yan, L. Lu, and M. Xu, "Anatomical invariance modeling and semantic alignment for self-supervised learniluong in 3d medical image analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 865–15 875. [Page 17.]

[38] J. Zhang, F. Zhan, C. Theobalt, and S. Lu, "Regularized vector quantization for tokenized image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 467–18 476. [Page 19.]

[39] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, and B. M. et al., "Monai: An open-source framework for deep learning in healthcare," 2022. [Online]. Available: https://arxiv.org/abs/2211.02701 [Page 32.]

[40] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: https://arxiv.org/abs/1711.05101 [Page 34.]

[41] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," 2018. [Online]. Available: https://arxiv.org/abs/1706.02677 [Page 34.]

[42] Y. Zhang and L. Liu, "Polynomial learning rate policy with warm restart for deep neural networks," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, p. 8929465. [Page 35.]

[43] R. Wightman, "Pytorch image models," https://github.com/rwightman/pytorch-image-models, 2019. [Page 35.]

[44] D. Müller, D. Hartmann, P. Meyer, F. Auer, I. Soto-Rey, and F. Kramer, "Miseval: a metric library for medical image segmentation evaluation," in *Challenges of Trustable AI and Added-Value on Health*. IOS Press, 2022, pp. 33–37. [Page 36.]