# Optimization of Pre-Operative Planning in Minimally Invasive Thoracic Surgeries with Deep Learning-based Patient-Specific 3D Modeling and Intuitive VR Interaction

**Arash Harirpoush**

**A Thesis**

**in**

**The Department**

**of**

**Computer Science and Software Engineering**

**Presented in Partial Fulfillment of the Requirements**

**for the Degree of**

**Master of Applied Science (Computer Science) at**

**Concordia University**

**Montréal, Québec, Canada**

**November 2024**

# CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Arash Harirpoush**

Entitled: **Optimization of Pre-Operative Planning in Minimally Invasive Thoracic Surgeries with Deep Learning-based Patient-Specific 3D Modeling and Intuitive VR Interaction**

and submitted in partial fulfillment of the requirements for the degree of

**Master of Applied Science (Computer Science)**

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

______________________________ Chair
*Dr. Tiberiu Popa*

______________________________ External Examiner
*Dr. Nizar Bouguila*

______________________________ Examiner
*Dr. Tiberiu Popa*

______________________________ Co-supervisor
*Dr. Marta Kersten-Oertel*

______________________________ Co-supervisor
*Dr. Yiming Xiao*

Approved by ______________________________
Dr. Joey Paquet, Chair
Department of Computer Science and Software Engineering

____________ 2024

______________________________
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

# Abstract

Optimization of Pre-Operative Planning in Minimally Invasive Thoracic Surgeries with Deep Learning-based Patient-Specific 3D Modeling and Intuitive VR Interaction

Arash Harirpoush

This thesis explores the application of deep learning algorithms within extended reality to enhance preoperative planning in minimally invasive video-assisted thoracic surgery (VATS). VATS faced technical challenges, such as a limited field of view and complex anatomical structures, which require precise, patient-specific 3D modeling and intuitive data interaction for effective planning. While deep learning, particularly U-shaped architectures, has emerged as a powerful approach for generating these models through automated segmentation of preoperative medical images, the growing number of U-shaped models with diverse network configurations and attention mechanisms requires systematic evaluation. Our first contribution addresses this need through a comprehensive benchmark study of U-shaped models, focusing on their segmentation accuracy and computational complexity. The study reveals the effectiveness of CNN-based U-shaped architectures for thoracic anatomical segmentation, with residual blocks playing a crucial role in enhancing network performance. These findings provide essential guidance for model selection and development in surgical planning applications, where the balance between accuracy and computational efficiency is important. Building upon these segmentation capabilities, our second contribution introduces an innovative extended reality system for optimizing trocar placement in VATS procedures. Optimal trocar placement is crucial to ensuring comprehensive thoracic cavity access, maintaining panoramic endoscopic visualization, and preventing instrument crowding. Our system features tailored visualization and interaction designs that enable surgeons to explore trocar configurations preoperatively using patient-specific 3D models. Preliminary evaluation demonstrates the system's efficiency, robustness, and user-friendliness, establishing its potential for clinical implementation while offering valuable insights for future surgical XR system development.

# Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisors, Dr. Marta Kersten-Oertel and Dr. Yiming Xiao, for their unwavering guidance and support throughout my journey. Their mentorship, insightful feedback, and innovative ideas have been invaluable, providing me with numerous opportunities to enhance my academic and research skills. I am honored to have had such incredible professors by my side. Your commitment to excellence and dedication, especially in providing individualized support and mentorship tailored to each student's needs, have been truly inspiring.

Besides my supervisors, I also extend my sincere appreciation to my wonderful colleagues in the APLab and HEALTH-X labs. Thank you for creating a supportive and collaborative environment where we could share knowledge and learn from each other.

I am especially grateful to Dr. George Rakovich for generously sharing his surgical expertise and invaluable insights. His guidance was instrumental in helping us define the requirements for our preoperative planning VR system. He patiently answered my questions, familiarized us with the challenges faced during surgery, and offered invaluable feedback that shaped the development of our system. I deeply appreciate his contributions to our research.

Finally, to my dear family and friends, thank you for your unwavering love and encouragement throughout this challenging journey. Your support was my constant source of strength, making this accomplishment possible.

Without the support and encouragement of these people, this journey would not have been possible. Thank you all.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Lung cancer is the second most common and deadliest cancer worldwide [79]. Surgical treatment often involves different types of resections depending on the disease's progression. A common procedure is lobectomy, where the lung lobe containing the tumor is removed [64]. Conventionally, this was performed through open surgery, requiring a large incision in the thorax to access the surgical site. However, with advancements in surgical techniques and technology, minimally invasive approaches like Robot-Assisted Thoracic Surgery (RATS) and Video-Assisted Thoracic Surgery (VATS) have gained popularity for early-stage lung cancer. These methods offer numerous benefits, including faster recovery, shorter operation time and hospital stay, and reduced blood loss.

In minimally invasive surgeries, small incisions, called trocars, are used to access and view the surgical area. These entry points must be selected strategically to ensure full access to the chest, allow the endoscope to provide a panoramic view, and prevent tool interference. Techniques like the Baseball Diamond Principle (BDP) and Triangle Target Principle (TTP) have been developed to achieve this. However, surgeons often rely on experience and patient-specific anatomical landmarks for trocar placement, which can lead to sub-optimal trocar placement, resulting in surgical complications and the need to relocate the trocars. Note that this relocation increases the invasiveness of the surgery for the patient and the operation time and could also increase the surgical team fatigue.

Preoperative planning with patient-specific 3D models and Extended Reality (XR) have shown promise in overcoming these challenges. For instance, Maddah et al. [56] reported up to a 40% reduction in procedure time for laparoscopic hysterectomy using a decision aid system, while López-Mir et al. [54] found a 33% improvement in trocar placement accuracy with Augmented Reality. These results showcased the potential of preoperative planning to improve precision and efficiency in minimally invasive surgeries. Inspired by this potential, we developed a patient-specific preoperative virtual system to investigate optimal trocar placement. This system requires accurate segmentation of relevant anatomy from preoperative images. Given the variety of deep learning models available for automated segmentation, we first benchmarked their performance in terms of accuracy and efficiency, considering clinical needs. Then, we developed a rule-based VR system to investigate optimized trocar placement within a patient's 3D model preoperatively. This system consists of two steps: first placing the endoscopic camera, followed by the surgical instrument trocars. The rules prioritize maximizing the reachable area for instruments, ensuring sufficient camera view, and preventing interference with critical anatomies.

## 1.1 Patient-Specific 3D Models in Surgical Planning

Integrating patient-specific models in surgical planning opens new avenues for enhancing procedural accuracy and safety. For complex anatomies like the pulmonary system, these models enable the surgical team to better understand tumor positioning relative to vital structures, aiding in precise diagnosis and reducing the risk of damaging critical arteries. This level of detail during surgical planning can minimize blood loss and improve surgical outcomes. In addition, patient-specific models also play valuable roles in medical education and training systems to enable the development of surgical skills without risking patients' well-being.

These patient-specific models are typically generated from preoperative imaging, such as CT or MRI scans. Traditionally, experts manually identified and segmented anatomical structures, making it time-consuming and costly. Advances in deep learning have introduced algorithms that can automate this process, accelerating model generation and allowing 3D models to be created more quickly and efficiently. This automation enhances accessibility, integrating patient-specific surgical planning tools into clinical workflows. For example, systems aimed at optimizing trocar placement powered by deep learning-driven segmentation can be integrated into surgical environments. Deploying the systems in the cloud can further extend accessibility by allowing them to be accessed via web servers, making them available to more clinical teams regardless of location. However, deep learning models used in these systems must balance high accuracy with efficient computational resource usage to ensure reliability.

## 1.2 Virtual Reality Systems in Surgical Planning

Recent studies highlight the promising potential of Virtual Reality (VR) environments in surgical procedures. These environments are types of Extended Reality systems providing fully immersive environments with virtual elements. In minimally invasive surgeries, VR has proven beneficial for training and preoperative planning, helping reduce the learning curve [43] in medical procedures and enhancing surgical preparation. By integrating patient-specific 3D models, VR environments allow surgical teams to explore detailed, personalized anatomy, supporting them in validating diagnoses and practicing various aspects of the procedure.

## 1.3 Contributions

This thesis advances preoperative planning for thoracic surgery by incorporating patient-specific models into a VR environment. Achieving this integration requires accurate segmentation algorithms to generate accurate 3D models. To automate this process, we examined various U-shaped deep learning architectures, which have demonstrated strong performance in medical imaging. Although many variants of these architectures exist, few studies explore how they affect 3D image analysis. In Chapter 3, we address this gap by investigating the accuracy and efficiency of different U-shaped model variants, focusing on the effect of attention mechanisms and architectural configuration on model performance and computational demands. These findings can guide the selection of optimal DL models for generating patient-specific segmentation maps.

Building on these segmentation maps, our second contribution, presented in Chapter 4, introduces a VR-based preoperative planning system to optimize trocar placement for thoracic surgeries.

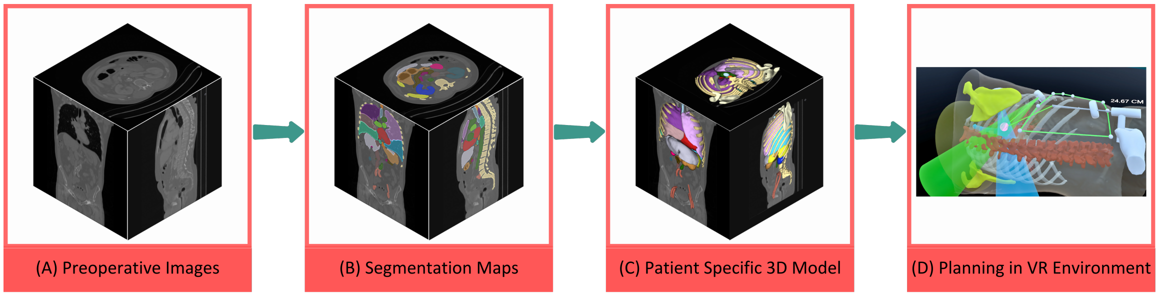


Figure 1.1: Overview of the proposed workflow: (A) Preoperative imaging of the patient, (B) Segmentation of torso anatomies related to the thoracic surgery, (C) Creation of patient-specific 3D models from segmentation maps, (D) Employing these 3D models in a VR environment to optimize trocar placement.

This system, designed for use with a head-mounted display (HMD), represents a pioneering approach to optimizing trocar placement in a VR environment. We demonstrate its functionality using right upper lobectomy, a common lung resection, where trocar placement consists of two steps: first placing the instrument trocars followed by camera placement for a panoramic view. The system integrates task-specific visualization, intuitive interaction mechanisms, and real-time user feedback to assist surgical planning. A preliminary user study with 20 participants evaluated the system's usability and robustness, indicating strong potential for its clinical application.

As a summary, Fig 1.1 shows the workflow, where Figs 1.1A and 1.1B represent the segmentation process discussed in Chapter 3, while Figs1.1C and 1.1D demonstrate the VR-based planning and optimization described in Chapter 4.

## 1.4 Outline

The remainder of this thesis is organized as follows. Chapter 2 provides the background and literature review related to our contributions and is divided into two main sections. The first section introduces various deep learning architectures, including the nnU-Net framework, along with a review of benchmark studies and their challenges. This section concludes with a brief overview of how Convolutional Neural Networks (CNNs) and Transformers perform in the frequency domain. The second section of Chapter 2 discusses foundational concepts of Extended Reality (XR), including the Reality-Virtuality Continuum, and presents a literature review of XR applications in surgical settings. In Chapter 3, we present our first contribution, a systematic benchmark study of U-shaped deep learning models for 3D medical image segmentation. Then, Chapter 4 details our second contribution, where we introduce the first preoperative HMD VR system designed to optimize trocar placement for thoracic surgeries. Finally, Chapter 5 concludes the thesis by summarizing our findings and discussing potential avenues for future research to build upon our contributions.

# Chapter 2

# Background

## 2.1 Deep Learning in Medical Image Segmentation

Deep learning has significantly impacted surgical planning by enabling the precise segmentation of anatomical structures from medical images. This facilitates the creation of patient-specific 3D models, giving surgeons a deeper understanding of the patient's anatomy. These models can then be integrated into extended reality (XR) environments, enhancing preoperative and intraoperative planning. U-shaped architectures, known for their effectiveness in medical image segmentation, have become a dominant approach due to their elegant design and strong performance. In this section, we explore the key components of these U-shaped models, providing an overview of their architectural designs and the associated functionalities.

### 2.1.1 CNN Auto-Encoder Models

Convolutional Neural Networks (CNNs) form the backbone of many medical segmentation models due to their ability to extract spatial features from images [52]. CNNs achieve this through convolutional filters that learn to recognize patterns in local image regions. These filters act like high-pass filters [62], making CNNs adept at capturing fine details and boundaries. CNN's typically integrated within an Auto-Encoder architecture to produce detailed segmentation masks that precisely delineate the target structures. As shown in Fig. 2.1, this architecture consists of two main components: an "Encoder" which extracts the features by leveraging CNN models, and a

“Decoder” which employs operations like transpose convolution or interpolation to upsample these feature maps and generates the final segmentation maps.

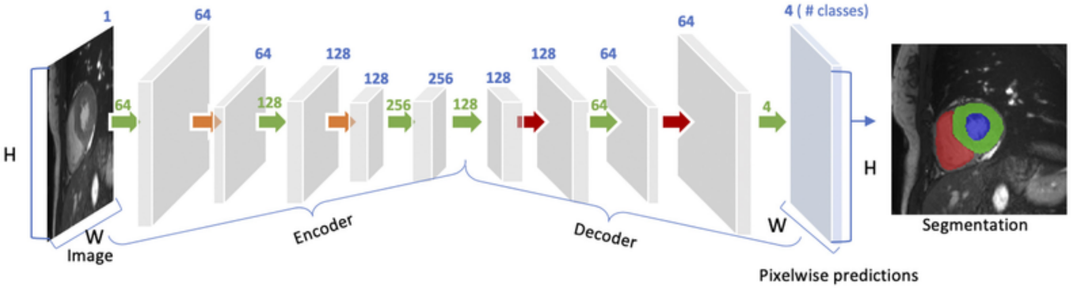
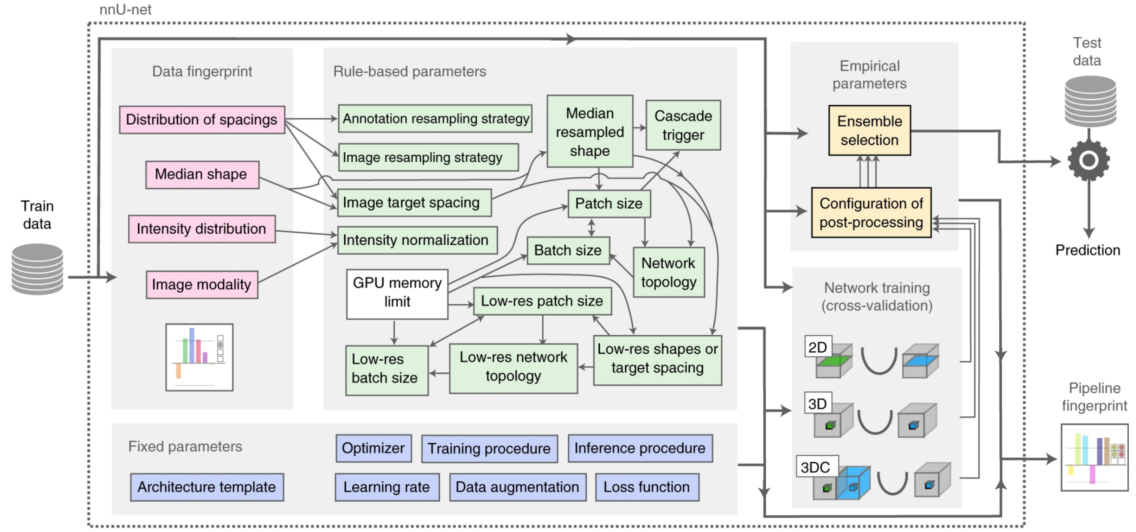


Figure 2.1: Overview of the CNN model for image segmentation [14]

### 2.1.2 nnU-Net framework

Developing deep learning models for medical image segmentation is complex, often requiring significant expertise and experience [40]. The nnU-Net framework simplifies this process by automating the design and training of 2D and 3D U-Net-based models, tailoring the architecture and hyperparameters based on the dataset and hardware specifications. To manage design choices efficiently, nnU-Net organizes parameters into three groups. Fixed parameters are those that don’t require task-specific adjustments, such as the use of Dice and Cross-Entropy loss functions, which are set according to standard practices. Rule-based parameters are guided by heuristic rules that link specific dataset features to design choices. For example, the initial patch size is set to the median image shape and then adjusted iteratively, alongside network topology changes, to allow training with a batch size of two on the available GPU memory. Lastly, empirical parameters are those refined through empirical tuning, including nuanced model architectural choices and post-processing techniques. The framework trains three U-Net configurations (e.g., 2D and 3D models) and selects the best-performing model, which may be a single model or an ensemble that combines softmax probabilities from two configurations by averaging them. This automated pipeline, as shown in Figure 2.2, enables nnU-Net to consistently achieve state-of-the-art performance in a wide range of medical segmentation tasks, often outperforming its counterparts [39].



| Design choice | Required input | Automated (fixed, rule-based or empirical) configuration derived by distilling expert knowledge (more details in online methods) |
|---|---|---|
| Learning rate | – | Poly learning rate schedule (initial, 0.01) |
| Loss function | – | Dice and cross-entropy |
| Architecture template | – | Encoder–decoder with skip-connection ('U-Net-like') and instance normalization, leaky ReLU, deep super-vision (topology-adapted in inferred parameters) |
| Optimizer | – | SGD with Nesterov momentum ($\mu = 0.99$) |
| Data augmentation | – | Rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma correction and mirroring |
| Training procedure | – | 1,000 epochs × 250 minibatches, foreground oversampling |
| Inference procedure | – | Sliding window with half-patch size overlap, Gaussian patch center weighting |
| Intensity normalization | Modality, intensity distribution | If CT, global dataset percentile clipping & $z$ score with global foreground mean and s.d.<br>Otherwise, $z$ score with per image mean and s.d. |
| Image resampling strategy | Distribution of spacings | If anisotropic, in-plane with third-order spline, out-of-plane with nearest neighbor<br>Otherwise, third-order spline |
| Annotation resampling strategy | Distribution of spacings | Convert to one-hot encoding →<br>If anisotropic, in-plane with linear interpolation, out-of-plane with nearest neighbor<br>Otherwise, linear interpolation |
| Image target spacing | Distribution of spacings | If anisotropic, lowest resolution axis tenth percentile, other axes median.<br>Otherwise, median spacing for each axis.<br>(computed based on spacings found in training cases) |
| Network topology, patch size, batch size | Median resampled shape, target spacing, GPU memory limit | Initialize the patch size to median image shape and iteratively reduce it while adapting the network topology accordingly until the network can be trained with a batch size of at least 2 given GPU memory constraints. for details see online methods. |
| Trigger of 3D U-Net cascade | Median resampled image size, patch size | Yes, if patch size of the 3D full resolution U-Net covers less than 12.5% of the median resampled image shape |
| Configuration of low-resolution 3D U-Net | Low-res target spacing or image shapes, GPU memory limit | Iteratively increase target spacing while reconfiguring patch size, network topology and batch size (as described above) until the configured patch size covers 25% of the median image shape. For details, see online methods. |
| Configuration of post-processing | Full set of training data and annotations | Treating all foreground classes as one; does all-but-largest-component-suppression increase cross-validation performance?<br>Yes, apply; reiterate for individual classes<br>No, do not apply; reiterate for individual foreground classes |
| Ensemble selection | Full set of training data and annotations | From 2D U-Net, 3D U-Net or 3D cascade, choose the best model (or combination of two) according to cross-validation performance |

Figure 2.2: Visual overview of the pipeline used in nnU-Net framework [40]

### 2.1.3 Vision Transformer

While CNNs have excelled in computer vision tasks, their limited ability to capture long-range dependencies led to the development of attention mechanisms. These mechanisms focus on different parts of an image to better capture long-range features. Among them, self-attention [84] assigns importance to various sequence tokens (e.g., image patches) and aggregates them accordingly. Multi-Head Self-Attention (MSA) extends this by applying multiple self-attention operations in parallel and then concatenating the results. Vision Transformers (ViTs) [19] use MSA to extract long-range features from input data. As shown in Figure 2.3.a, the input image is divided into 16x16 patches, which are processed and aggregated through MSA within Transformer blocks (Figure 2.3.b). However, despite their strong performance, ViTs suffer from weak local inductive bias and quadratic computational complexity, which led to the development of the Swin Transformer. This architecture introduces a window-based MSA (WMSA), which limits attention computations to local non-overlapping windows while connecting information across windows (Figure 2.3c) [53]. This approach not only strengthens local inductive bias, but also reduces computational complexity from quadratic to linear. As illustrated in Figure 2.3d, Swin Transformer blocks use WMSA, allowing for smaller patch sizes (e.g., 4x4), enabling the capture of finer image details in comparison to conventional ViTs.

### 2.1.4 Focal Modulation

To further enhance long-range feature extraction in vision models, Yang et al. [92] introduced "Focal Modulation" (Figure 2.4a) as an alternative to self-attention (SA). While SA computes pairwise interactions between all tokens, Focal Modulation generates an attention map through a "Context Aggregation" process (Figure 2.4b). This process uses a series of depth-wise convolutions with varying kernel sizes to capture multi-scale features, which are then selectively aggregated by a "Gated Aggregation" module based on their importance. The resulting attention map is applied to the input features through element-wise or affine transformations, effectively modulating the input based on the learned context.

### 2.1.5 Benchmarking U-Shaped Architectures in Medical Image Segmentation

The availability of public datasets like BTCV and BraTS has enabled multiple studies to benchmark U-shaped models for medical image segmentation. Early research by Gut et al. [25] and Kugelman et al. [46] used the nnU-Net framework to evaluate these models in a standardized

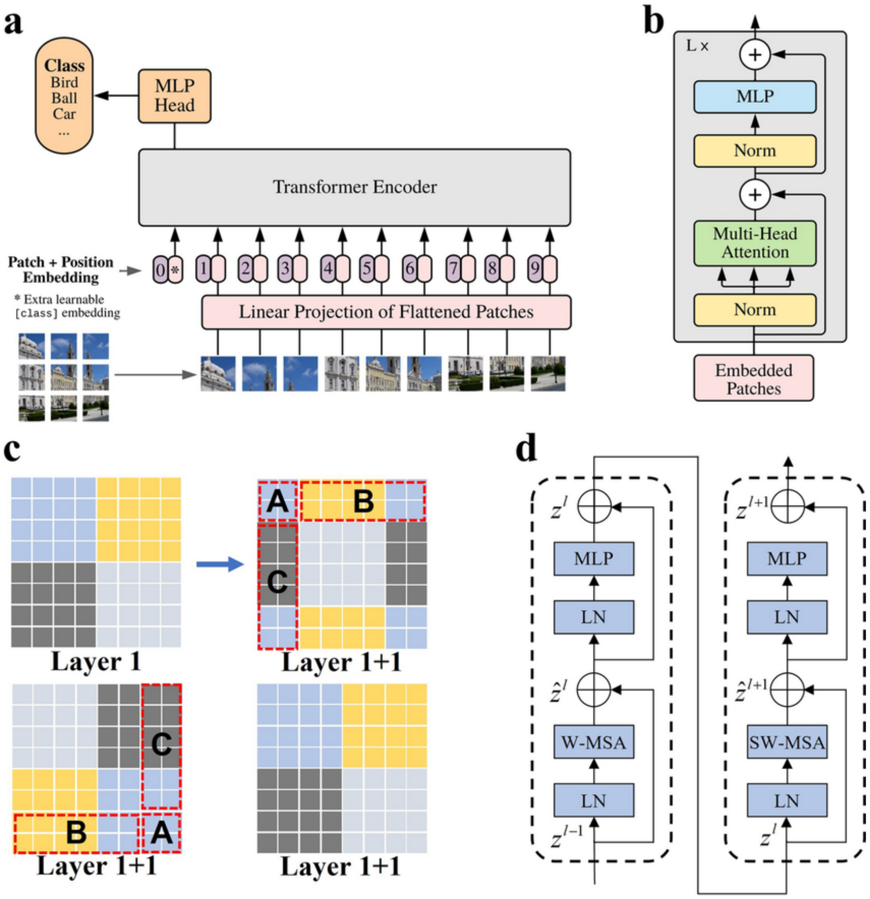


Figure 2.3: Overview of the Vision and Swin Transformer models. (a) The ViT architecture transforms input feature maps into patches, applies linear mapping, and processes them through the Transformer, with final classification handled by an MLP. (b) Details of the ViT encoder emphasize the role of multi-head attention modules in feature extraction. (c) In the Swin Transformer, feature maps evolve through Window-based Multi-Head Self-Attention (W-MSA) and Shifted Window Multi-Head Self-Attention (SW-MSA), with a cyclic shift operation that enables feature integration across shifted windows. (d) The Swin Transformer Block summarizes this process, outlining key computational steps [66].

pipeline. Their findings consistently showed that increasing architectural complexity did not necessarily lead to better performance, with the basic U-Net often outperforming more advanced variants across various tasks and datasets. As architectures evolved, Vision Transformers were introduced into U-shaped models to improve performance by capturing long-range dependencies. However, their effectiveness remains debatable. Many studies [44, 93, 90] agree that fully Transformer-based models struggle due to a lack of local inductive bias and increased computational overhead, which hinder performance. Some research, including Xiao et al. [90] and Yao et al. [93], suggested that

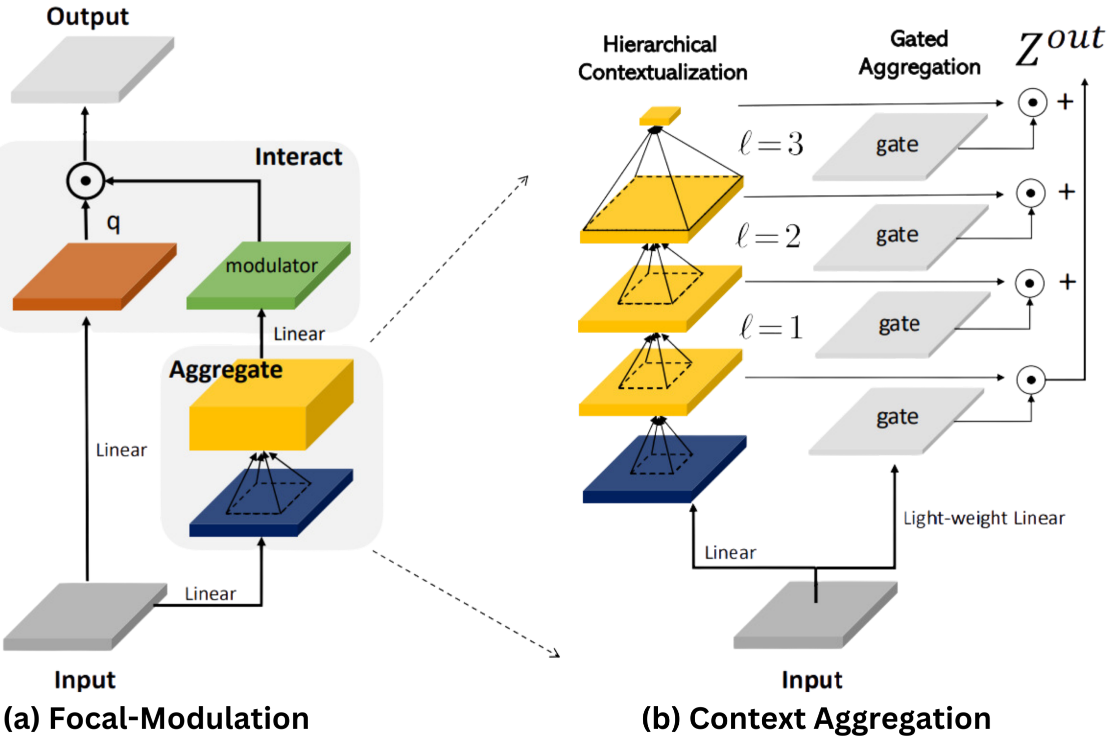


Figure 2.4: Overview of the Focal Modulation. (a) Focal Modulation block. (b) Detailed visualization of "Context Aggregation" in Focal Modulation [92].

incorporating Transformers into CNN architectures can be beneficial while others, such as Ji et al. [44], found no significant performance gains. For example, the study of Ji et al. [44] on CT and MRI images showed that the classic U-Net consistently outperformed hybrid and fully Transformer-based models.

Addressing the inconsistencies in previous benchmarks, Isensee et al. [41] highlighted common pitfalls in model comparisons and recommended using the nnU-Net [40] framework to ensure fair evaluations. Their comprehensive benchmarking of CNN, Transformer, and Mamba-based U-shaped models validate earlier conclusions, demonstrating that increasing architectural complexity does not always translate into improved performance, further confirming the observations of Gut et al. [25] and Kugelman et al. [46].

### 2.1.6 CNNs and Transformers in the Frequency Domain

Park and Kim [62] analyzed the behavior of CNNs and Transformers in the frequency domain to better understand their complementary strengths. They found that while CNNs act as high-pass filters, capturing high-frequency details, multi-head self-attention (MSA) mechanisms in Transformers serve as low-pass filters, focusing on low-frequency information. This combination of filtering behaviors suggests that integrating the two architectures can improve performance. Models like TransUNet [15] have already demonstrated the effectiveness of applying transformers after CNN layers, leading to better segmentation results. Although this sequential approach was initially adopted due to computational limitations, recent studies, such as the SwinUNETRV2 [33], indicate that strategically placing residual blocks before Transformer layers can further boost performance. Supporting these findings, Park and Kim [62] visualized ViT's feature map, shown in Figure 2.5. Their analysis in the Fourier domain showed that MSAs in the early stages amplify high-frequency signals, similar to CNNs. This amplification helps explain why hybrid models, where Transformers follow CNN layers, deliver outstanding performance compared to fully-transformer-based models by combining the strengths of both architectures. However, as mentioned previously in subsection 2.1.5 these enhancements could not necessarily be due to the architectural differences. In most cases, the limited amount of available data resulted in better performance of CNN models.

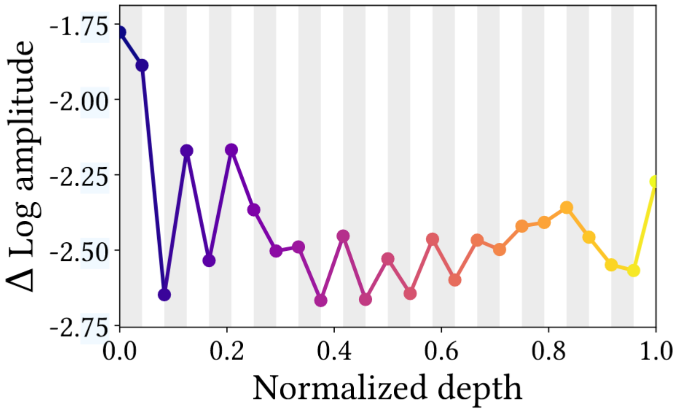


Figure 2.5: The Δ Log amplitude of MSA at high frequency (1.0 $\pi$). Gray regions show a reduction of high-frequency features by MSA, while white regions display amplification by Multi-Layer Perceptron (MLP) layers [62].

## 2.2 Extended Reality (XR) Technologies for Surgical Applications

Extended Reality (XR) contains a spectrum of immersive technologies that seamlessly merge real and virtual elements, creating novel environments and experiences with the potential to revolutionize various fields, including medicine. To introduce the diverse landscape of XR technologies, we turn to the Milgram and Kishino [59] Reality-Virtuality (RV) Continuum (see Figure 2.6), which classifies different technologies based on how much they mix the real and virtual components. Towards the "real" end of the continuum lies Augmented Reality (AR), where digital content is overlaid onto the real world, enhancing the user's perception of their surroundings with supplementary information. In surgical applications, this could involve overlaying patient-specific anatomical data onto the surgeon's field of view during a procedure.

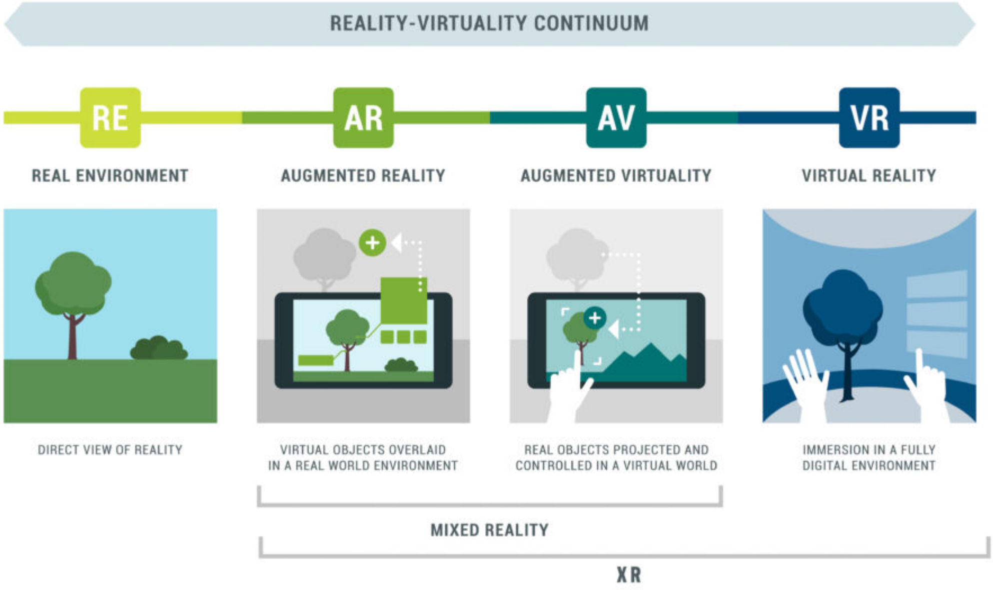

REALITY-VIRTUALITY CONTINUUM
RE
AR
AV
VR
REAL ENVIRONMENT
AUGMENTED REALITY
AUGMENTED VIRTUALITY
VIRTUAL REALITY
DIRECT VIEW OF REALITY
VIRTUAL OBJECTS OVERLAID IN A REAL WORLD ENVIRONMENT
REAL OBJECTS PROJECTED AND CONTROLLED IN A VIRTUAL WORLD
IMMERSION IN A FULLY DIGITAL ENVIRONMENT
MIXED REALITY
XR

Figure 2.6: Overview of the Milgram and Kishino [59] Reality-Virtuality (RV) Continuum, showing the spectrum of immersive experiences from the real world to virtual environments [17]

Moving further along the spectrum, we encounter Augmented Virtuality (AV), which introduces real world elements into predominantly virtual environments. This can be seen in a camera placement system where a physical device allows users to manipulate the camera within a virtual environment to explore the entry point, providing haptic feedback and displaying a physical representation of the device in the virtual environment.

Milgram and Kishino [59] grouped AR and AV under the term Mixed Reality (MR), defining it as any environment, where real and virtual objects coexist. However, recent research by Skarbez et al. [76] expands this definition. They suggest that MR should be seen as environments, where real and virtual elements not only coexist but also interact dynamically. In the surgical context, a virtual model can be augmented on a physical phantom to simulate surgical procedures like organ resection, combining real-world interaction with detailed virtual visualization for enhanced training.

At the “virtual” end of the continuum lies Virtual Reality (VR), where users are fully immersed in a digital environment [59]. While VR typically engages the user’s exteroceptive senses, such as perceiving external stimuli like sight and sound, it cannot fully control their interoceptive senses, which provide awareness of the body’s internal state, such as balance and spatial orientation [76]. This persistent connection to the real world through interoception places VR within the broader Mixed Reality spectrum, despite its primary focus on creating virtual experiences [76]. In surgical training, VR can provide a safe and controlled environment for practicing procedures without the risk of harming real patients.

### 2.2.1 XR for Surgical Path Planning

Several studies have highlighted the benefits of preoperative path planning systems across various surgical interventions. In minimally invasive lateral skull-based surgeries, single-port approaches were traditionally used. However, multi-port approaches offer advantages, such as enhanced instrument manipulation, direct visual feedback, and broader applications like tumor removal. Multi-port setups also provide more space around the surgical area where instrument trajectories intersect [78]. Stenin et al. [78] investigated the feasibility of multi-port approaches of lateral skull base surgery, like tumor removal and biopsy, using a planning tool that mapped trajectories inside patient-specific 3D temporal bone models created from CT scans. They plan three ports: one for an endoscopic camera and two for surgical instruments, focusing on maximizing the distance between critical structures and optimizing the angle between instrument paths. Similar to minimally invasive thoracic surgery [73], Stenin et al. [78] emphasized that the procedure can be facilitated when the instruments meet each other at the appropriate angle near the surgical target.

The study stated that while intersecting instrument paths beyond the target allows for better manipulation space, intersections that converge too closely can lead to instrument conflicts. This can be analyzed by measuring the manipulation angle, the degree to which instruments meet each other, with higher angles indicating greater distance between the ports.

In a study by Schwenderling et al. [74], they introduced a projector-based augmented reality (AR) system to address challenges in identifying insertion points for percutaneous interventions. Surgical Planning like identifying the entry points in traditional 2D imaging can be time-consuming, and manually transferring insertion points to the skin increases the risk of error [74]. The AR system introduced by Schwenderling et al. [74] projected insertion and target points onto a torso phantom, allowing users to select the insertion point with a needle. Path quality was evaluated based on three factors: distance from critical anatomy, path length, and insertion angle. Hard conditions were used to exclude unsafe paths, while soft conditions were applied to score path quality. To explore the impact of different visualization techniques on the user's decision making they provided two visualizations for the insertion based on the path quality: "Area Visualization", in which all possible paths were projected onto the phantom, and "Full Visualization", which paths were color-coded based on their score. The study found that inexperienced users benefited from the full visualization technique, while the target visualization had minimal effect on their decision-making. These studies demonstrate the potential of XR systems to enhance surgical path planning and improve decision-making in complex procedures.

### 2.2.2 VR for Surgical Simulation

When it comes to surgical simulation, VR environments play a crucial role in education and preoperative practice based on patient anatomy. Studies have focused on the applicability of these systems and identifying metrics to evaluate procedure quality and differentiate participants by their experience level. Early VATS training relied on animal models, which lacked anatomical variation, leading to the development of XR environments for surgical simulation. To address this issue, Solomon et al. [77] introduced a VR environment on a personal computer (PC), providing a 3D view of anatomical structures from the perspective of an endoscope camera and an external view. It also used haptic devices to control instruments simulating video-assisted thoracoscopic surgery (VATS)

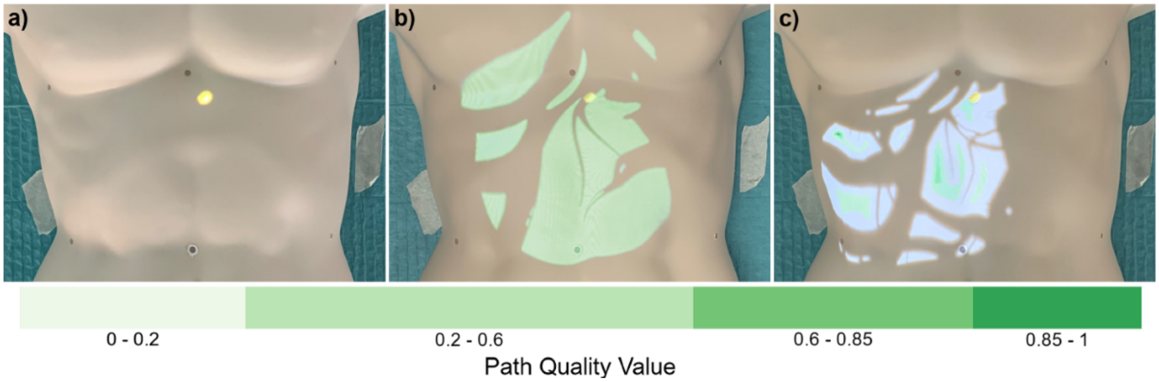


Figure 2.7: The insertion visualization projected possible insertion points onto the skin that met the hard path planning conditions. In the baseline view (a), no insertion points were displayed. In the Area visualization (b), there was no additional information on path quality. In the Full visualization (c), the insertion points were color-coded based on their associated path quality values, which were determined using soft path planning conditions. A green color scale was employed, where darker shades indicated better paths with higher path quality values [74].

for a right upper lobectomy, demonstrating the feasibility of such simulations. The system allowed users to select port locations via a mouse and manipulate surgical tools with haptic handles. For the system, the first step involves identifying the endoscopic port and visualizing its view on a monitor. As users explore the chest cavity, they identify other ports. Notably, the endoscopic port could be adjusted during the procedure based on the locations of the other ports. The proposed system could detect common issues, such as pulmonary parenchyma tears. While this proposed system advanced surgical training methods, it still lacked surgical performance scoring, the incorporation of patient-specific models, and physics-based tissue responses.

In order to identify performance metrics and refine VR systems, Jensen et al. [43] developed a VR system inspired by the LapSim simulator, specifically designed for right upper lobectomy. Their system incorporated various performance metrics, such as total procedural time, blood loss, and hand movements, and was tested by over 100 thoracic surgeons with varying levels of experience. The aim was to develop metrics that could differentiate participants' skill levels while assessing the system's effectiveness from the surgeons' perspective. The simulation followed a standard three-port approach and included anatomical dissections and stapling. Four fixed endoscopic views were provided, though the system also allowed assistance to control the camera. Despite the promising feedback from experienced surgeons, who found the simulation realistic and beneficial for novice and intermediate surgeons, no clear metrics appeared to distinguish participants by experience due to the complexity of the procedure and varying techniques. The study of Jensen et al. [43] emphasized realistic simulation through enhanced graphics and physics-based tissue responses, but the challenge of identifying experience-based metrics and incorporating patient-specific models remained.

Haidari et al. [26] expanded on this work by developing a VATS lobectomy module for the LapSim VR simulator, covering all five lobes. To validate the system and explore metrics for evaluating surgical quality, 45 surgeons with different experience levels (novice, intermediate, and experienced) participated in the study, similar to the study of Jensen et al. [43] design. Each participant performed three lobectomies within the system, and their performance was evaluated based on nine predefined metrics, including instrument path length and total procedural time. The study successfully identified three metrics that could differentiate the quality of results based on experience: mean procedure time, mean blood loss, and total instrument path length. While this study demonstrated the feasibility of establishing pass/fail levels concerning the quality of the surgical skills for simulation results based on contrasting groups, the high cost of the LapSim VR simulator (reported as 79,000 €) limited its accessibility.

Beyond training, VR simulations have broader applications. In a study by Ujiie et al. [83], the potential of an HMD-based VR surgical navigation system was explored for use in RATS (robot-assisted thoracic surgery). This system employed volumetric, patient-specific 3D reconstructions generated from contrast-enhanced CT scans. Perspectus VR education software was then used to interact with these models preoperatively, allowing for actions such as rotation and cropping to better understand the patient's anatomy. These models were further integrated with the surgical console's endoscopic video feed using the TilePro multi-display platform, allowing surgeons to view multiple information sources simultaneously. Ujiie et al. [83]'s findings highlighted the benefits of patient-specific models in a VR environment, enhancing a surgeon's understanding of pulmonary anatomy and potentially improving the accuracy and safety of RATS through better surgical planning.

### 2.2.3 Trocar Placement Planning within XR

In robot-assisted minimally invasive surgeries, proper trocar placement is essential for accessing the surgical area and avoiding collisions between robotic arms, which can compromise surgical

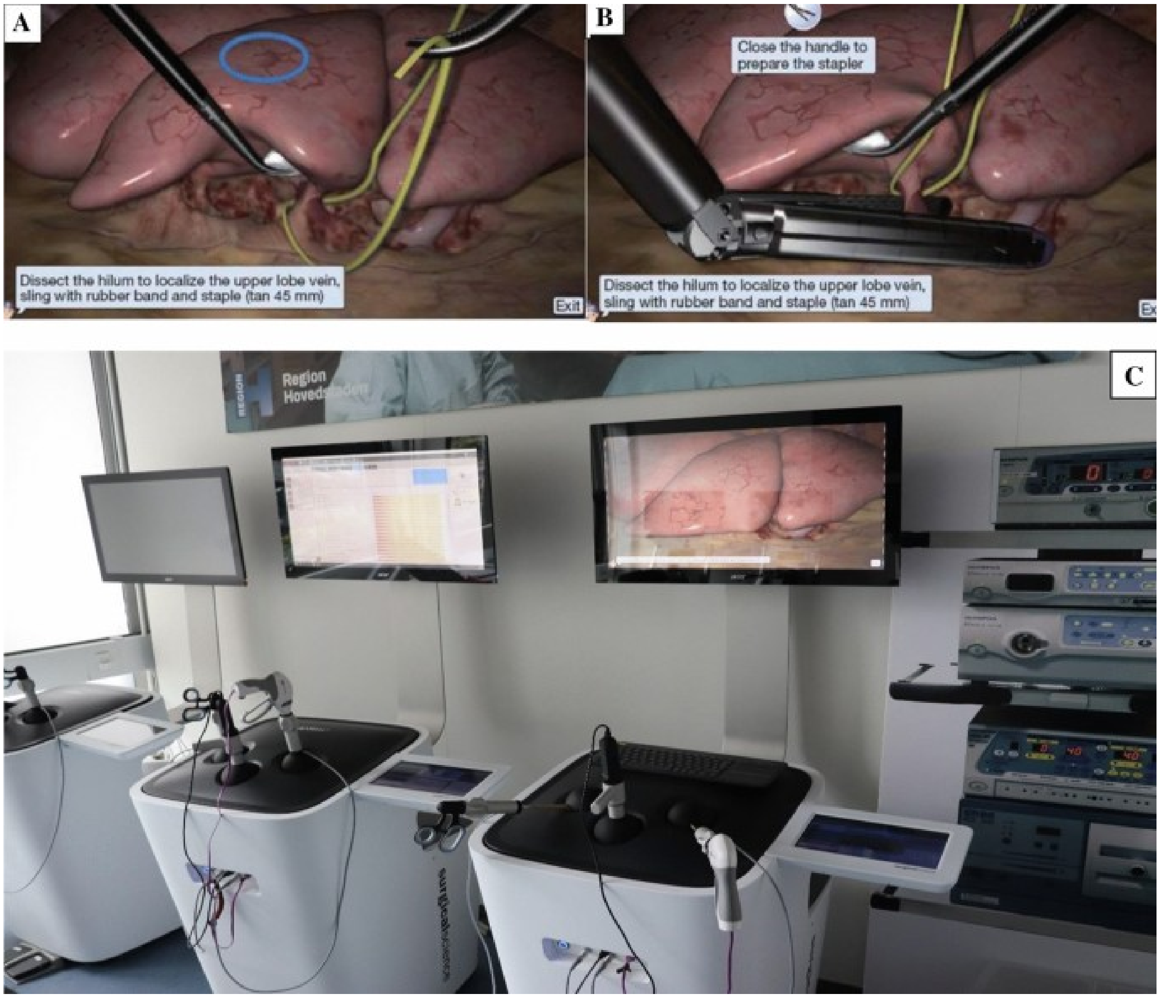

A
Dissect the hilum to localize the upper lobe vein, sling with rubber band and staple (tan 45 mm)
Exit
B
Close the handle to prepare the stapler
Dissect the hilum to localize the upper lobe vein, sling with rubber band and staple (tan 45 mm)
Ex
C
REGION
Region Hovedstaden
surgicalscience

Figure 2.8: VATS module on the virtual reality simulator LapSim. (A) Dissection and (B) stapling of the middle lobe vein on the VATS module on the LapSim virtual reality simulator. The blue circle shows the tumor location. (C) The virtual reality simulator setup used in Haidari et al. [26].

precision. Studies have shown that optimizing trocar placement can significantly reduce planning and procedural time. Simoes and Cao [75] developed a mixed reality system to help surgeons in trocar placement decisions. This system uses an optimization algorithm to calculate initial trocar locations based on patient anatomy and the surgical procedure in robot-assisted laparoscopic surgery. The trocar locations are then projected onto the patient's abdomen, allowing the surgeon to interact with them and adjust the placement. Similarly, Maddah et al. [56] explored the use of decision-aid systems for trocar placement in robotic-assisted hysterectomies. Their system also uses an optimization algorithm, incorporating patient-specific and robot models to determine optimal trocar placement. The system was tested on four patients, with two surgeons providing feedback on its performance. To test the system's usability, two surgeons performed simulated surgery on a torso phantom, completing the task with and without the decision-aid system across three target areas. The results showed a reduction of up to 40% in the total time needed to complete the given task when using the decision aid system, showcasing the potential benefits of preoperative planning for improved surgical outcomes and reduced procedure time.

The studies of Simoes and Cao [75] and Maddah et al. [56] provide compelling evidence for the potential benefits of preoperative planning tools in robot-assisted surgeries. Simoes and Cao [75] demonstrated the feasibility of using a mixed reality system to guide trocar placement in robot-assisted laparoscopic surgery. Meanwhile, Maddah et al. [56] found that a decision-aid system could significantly reduce procedural time in robotic-assisted hysterectomies. These findings highlight the crucial role of proper trocar placement in ensuring surgical precision and efficiency. Inspired by these findings, we investigated preoperative planning of trocar placement within a VR environment for Video-Assisted Thoracic Surgery. This immersive system allows surgeons to interactively explore and refine their approach before surgery, potentially enhancing surgical outcomes and reducing procedural time.

# Chapter 3

# Architecture Analysis and Benchmarking of 3D U-shaped Deep Learning Models for Thoracic Anatomical Segmentation

A version of this chapter has been published in *IEEE Access*:

- Harirpoush A, Rasoulian A, Kersten-Oertel M, Xiao Y. Architecture Analysis and Benchmarking of 3D U-shaped Deep Learning Models for Thoracic Anatomical Segmentation. *IEEE Access*,12, 127592 - 127603, 2024 [29].

## 3.1 Introduction

In modern surgical planning that emphasizes high precision and low trauma, 3D anatomical segmentation from pre-operative medical images is becoming increasingly important. Thoracic surgery, i.e., chest surgery which involves operations on lungs affected by cancer, trauma, pulmonary disease, or cardiac conditions, accounts for approximately 530,000 cases per year in the US [8]. In addition to video-based surgical guidance with limited spatial information, recent studies [67, 24] have demonstrated significant advantages of using patient-specific physical or digital 3D models for various thoracic surgeries [67, 24] in both conventional and mixed reality surgical environments [82, 12, 72]. To ensure the outcomes of these applications, efficient and accurate 3D anatomical segmentation and reconstruction is essential.

Deep learning (DL) approaches, such as convolutional neural networks (CNNs) have dominated the state-of-the-art performance in various radiological tasks. With their quick inference time, they offer a tool to enable efficient digital twin construction for thoracic surgical planning, simulation, and intra-operative monitoring. U-shaped models, pioneered by the 2D UNet [70], stand out among DL segmentation models [52] for their robust performance and elegant architecture. The typical U-shaped architecture comprises three key elements: an encoder for learning relevant image features and compressing them into lower-dimensional embeddings, a decoder for expanding these embeddings and producing the final segmentation, and skip connections that maintain fine-grained details during upsampling by aggregating feature maps across encoder and decoder layers. Since the first inception, major efforts have been dedicated to adapting the 2D framework to 3D [16], exploring different backbones for the encoder/decoder [40, 38, 61, 30, 68, 9], updating the resolution stages [94], and experimenting with novel network configurations [25]. To better understand the impact of these enhancements and investigate the application of 3D anatomical reconstruction for thoracic surgical planning and simulation, a comprehensive benchmark study, providing the models' architectural characteristics, would be highly instrumental, but has yet to be conducted.

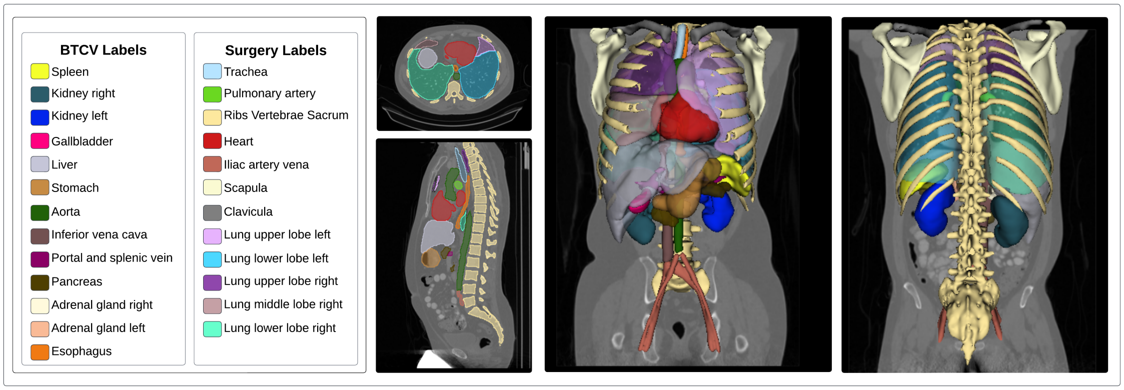


Figure 3.1: Demonstration of the anatomical structures for U-shaped model benchmarking, including 12 labels for thoracic surgery and 13 labels that are consistent with the BTCV segmentation challenge.

### 3.1.1 Variants of U-shaped Architectures

Many variants of the U-shaped architectures have been proposed primarily for medical image segmentation [70]. Among these, the nnUNet [40] framework, which allows task-specific optimization of 2D and 3D UNet models and training strategies, has achieved great success in a wide range of segmentation tasks. Recently, Huang *et al.* [38] proposed the STUNet, an enhancement of 3DUNet model from the nnUNet [40] framework, with modified downsampling and upsampling blocks in the encoder and decoder, respectively.

Besides architecture and training strategy optimization, some attempts have also been made to incorporate various attention mechanisms in U-shaped models. In the AttentionUNet [61], attention gates that combine trained soft attention feature maps through skip connections are implemented to enhance accuracy and model transparency for pancreas segmentation. With the emergence of the Vision Transformer (ViT), which leverages self-attention to capture long-range dependencies within an image, CNN-Transformer-hybrid U-shaped models were introduced to enhance the performance of the CNN-based UNet.

For example, the TransUNet [13], CoTR [91], and TranBTS [86] models strategically incorporated Transformer layers into the concluding stages of the encoder to enable enhanced features extraction from the feature maps generated by the preceding CNN blocks. More recently, the UNETR [31] and SwinUNETR [30] advocate for utilizing fully-transformer-based encoders by using ViT and Swin Tranformer [53], respectively, while retaining CNN decoders. Attempts have also been made to create U-shaped models driven fully by Transfomer blocks. Notably, the 2D Swin-UNet [9] first implemented this approach and demonstrated its performance on abdominal CT segmentation. Similarly, the VT-UNet [63] adopts the same approach, with further addition of cross-attention mechanisms in the decoder. In the same category, Zhou *et al.* [96] proposed the nnFormer, which introduced local and global self-attention mechanisms in the encoder, decoder, and skip-attentions. Lastly, the more recent Focal Modulation [92] offers an alternative attention mechanism that models hierarchical contextualization of image features, which are aggregated for each query token. This relatively new technique was shown to outperform its state-of-the-art counterparts, such as Swin and focal Transformers in 2D natural image recognition and segmentation tasks. Leveraging this new mechanism, the FocalSegNet [68] replaces the Transformer blocks of the encoder in the UNETR with new 3D Focal Modulation blocks to perform volumetric medical image segmentation.

Aside from attention mechanisms in U-shaped models, earlier research has explored various setups for the number of resolution stages and diverse skip connection schemes [97, 89, 85, 13] to improve segmentation accuracy. Among these, UNet++ [97] introduced nested hierarchical skip connections to fuse encoder and decoder features, and the BiO-Net [89] and its variants leveraged bidirectional skip connections within the U-Net model. Another technique in this context is to incorporate Transformers in skip connections for 2D medical image segmentation that is more computationally intensive [85, 13].

### 3.1.2 Previous Anatomical Segmentation Benchmarking studies

The increasing amount of public datasets, such as Kits [35], CHAOS [45], SegTHOR [47], BTCV [48], and ACDC [6] have greatly facilitated the benchmarking of various new algorithms in medical image segmentation. For 2D OCT image segmentation, Kugelman *et al.* [46] tested eight U-shaped models (VanillaUNet, DenseUNet, AttentionUNet, SEUNet, ResidualUNet, R2UNet, UNet++, and InceptionUNet), and concluded that the VanillaUNet was the most appropriate considering computational complexity while all models performed similarly. Yao *et al.* [93] benchmarked UNet, UNet++, TransUNet, and SwinUnet for 2D segmentation tasks of the lung and abdominopelvic ovarian masses, and they showed that the TransUNet model had the best performance in both datasets. In their paper, Ji *et al.* introduce the AMOS [44] MRI and CT dataset for 3D abdominal anatomy segmentation and tested UNet, VNet [60], CoTr, UNETR, SwinUNETR, and nnFormer on the proposed dataset. When considering both accuracy and model complexity (i.e., model parameters and the number of GFlops), they found no advantage of using transformers over CNNs, with the UNet achieving the best accuracy, particularly for larger anatomies. After reviewing various U-shaped transformer-based models in 2D and 3D medical image segmentation, Xiao *et al.* [90] evaluated Swin-Unet, LeViT-UNet, UCTransNet, TransAttUnet, UTNet, and UTNetV2 on MSD [1] and NSCLC [3] datasets for 2D segmentation of pancreas and lung cancer, respectively. Their study showed that while UTNetV2 is the best for pancreas segmentation, TransUNet achieved the best performance in lung cancer segmentation. They also found that combining CNN and Transformer was beneficial, and the choice of the best model can be task-specific.

### 3.1.3 Novelty and Contributions

While several groups have benchmarked U-shaped models as systematic studies or validation of a new algorithm, there are still remaining research gaps that we aim to address in this study. First, few studies investigated 3D segmentation of the lung [90] and other related anatomies for thoracic surgical planning and simulation. Second, many studies focused on 2D segmentation tasks and models while direct 3D processing can offer better spatial continuity across slices in the segmented labels. Finally, there is a lack of dedicated exploration for the impact of various attention mechanisms and network configuration designs on multi-organ 3D segmentation tasks. To tackle these, our study evaluates the performance of several U-shaped models, including 3DUNet, STUNet, 3D AttentionUNet, SwinUNETR, FocalSegNet, and a new 3DSwinUnet along with its variants, in segmenting anatomical structures associated with thoracic surgery from CT scans of the TotalSegmentator dataset [87]. To allow an easy comparison of conclusions with concurrent literature, we also benchmarked these models for the anatomies in the BTCV [48] challenge in the same dataset (see Fig. 3.1). Our main contributions can be summarized as follows:

- Investigate the effects of different attention mechanisms (attention-gate, self-attention, focal modulation, and baseline UNet) in U-shaped architectures

- Benchmark state-of-the-art U-shaped models, including our novel 3D SwinUNet and its eight variants in 3D image segmentation for thoracic surgery and BTCV challenge anatomies by considering their accuracy and computational complexity
- Investigate the impacts of different network configuration designs, including the number of resolution stages (number of upsampling/downsampling operations), different operations for skip connections, downsampling, and upsampling, and the effect of the bottleneck operation in a pure Swin Transformer-based 3D U-shaped model

The nuanced insights from this study can directly contribute to the development of more advanced and robust segmentation algorithms for medical image analysis and thoracic surgical planning. Specifically, by well understanding the strengths and limitations of different model architectures and operations, our study can better inform future DL model designs that balance accuracy and efficiency to meet the needs of different clinical contexts. The patient-specific 3D models derived from these enhanced segmentation techniques can offer promising avenues for improving surgical precision and patient outcomes. For instance, integrating these models into surgical virtual reality can offer fast and accurate patient-specific digital twin creation to allow safer and more efficient surgical planning, such as more accurate tumor localization [2] and optimized trocar placement [65]. These would ensure better treatment outcomes and time-saving in the often resource-limited clinical environment.

## 3.2 Methods and Materials

### 3.2.1 Dataset

We employed the dataset from the TotalSegmentator paper [87], which included 1204 body CT scans with 104 different labels. These labels covered 27 organs, 59 bones, 10 muscles, and 8 vessels. For our study, we focused on 79 annotations that are relevant to thoracic surgery and correspond to the BTCV challenge. As there is no need for further division of some anatomies and to facilitate training and evaluation, we combined some of the annotations that belong to the same anatomical structure, resulting in 25 labels: 12 for thoracic surgery and 13 that correspond to the BTCV challenge. The list and visualization of these anatomies are shown in Fig. 3.1. As the CT scans in the dataset contain different fields-of-view and an inconsistent number of labels, we selected the cases that contain more than 22 of the 25 classes that we defined, resulting in 440 training cases, 26 validation cases, and 30 testing cases. To facilitate the assessment, we also provide the distributions of each anatomical label across the training and testing datasets in Fig. S1 of the *Supplementary Materials*.

### 3.2.2 Experimental set-up and network training

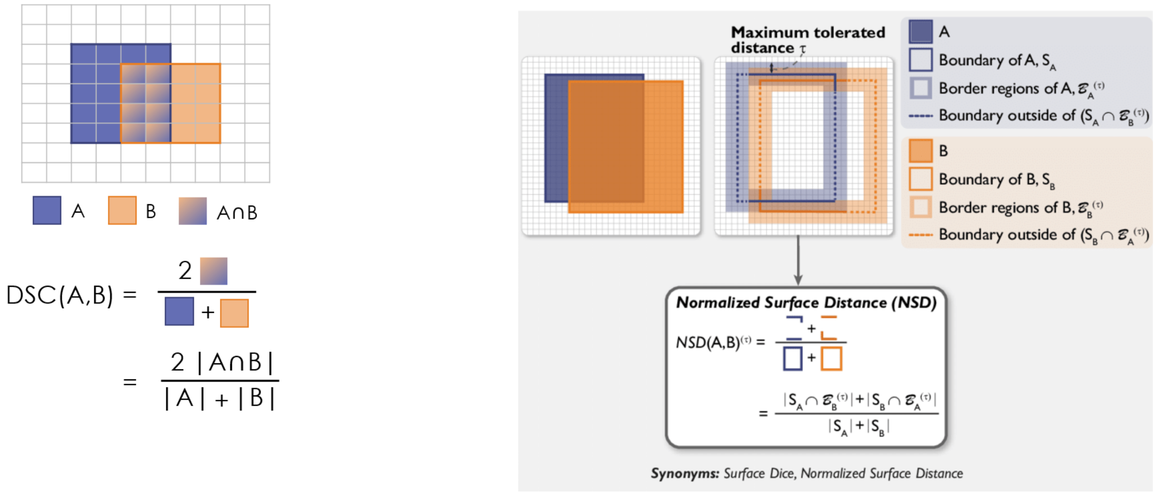
We selected six U-shaped models for performing the benchmarking, including a 3DUNet optimized based on the nnUNet framework [40], STUNet [38], AttentionUNet [61], Swin-UNETR [30], FocalSegNet [68], and a new 3D implementation of the original 2D Swin-Unet [9] (we refer to as 3DSwinUnet). A visual summary of these models' architectures is included in Fig. S2 of the *Supplementary Materials*. Furthermore, we also devised 8 additional variants of the 3DSwinUnet with different strategies for skip connections, feature map downsampling, upsampling, and bottleneck (more details in Section 3.2.3). Such a range of models allows us to comprehensively assess the impact of different attention mechanisms, as well as more nuanced network configuration designs for U-shaped models. Specifically, the STUNet [38] and FocalSegNet [68] models were taken from their github repositories while implementation of 3D AttentionUNet and SwinUNETR models in MONAI [10] were used. Finally, the 3DSwinUnet and its variants were implemented using the MONAI framework. The 3DUNet, STUNet, and AttentionUNet models contain five resolution stages while the SwinUNETR and FocalSegNet models have four resolution stages. Finally, the 3DSwinUnet model and its variants consist of three resolution stages [9].

As the dimension of the input image patch is $96 \times 96 \times 96$ voxels across all selected neural network architectures, to allow a similar field-of-view per 3D input patch to the original TotalSegmentator paper [87] for comparison, we resampled the CT scans to a $2.0 \times 2.0 \times 2.0mm^3$ resolution. For all model training, we used the SGD optimizer with an initial learning rate of 0.01, a Nesterov momentum of 0.99, and a weight decay of $1e-3$ to minimize the loss function which is a sum of the cross-entropy and Dice loss. For the learning rate scheduler, we employed the Poly method, which reduces the learning rate by increasing the number of epochs by calculating

$LR \times (1 - Epoch/MaxEpoch)^{0.9}$, where $LR$ represents the initial learning rate, $Epoch$ is the current epoch number, and $MaxEpoch$ is the maximum epoch number. Each epoch consists of 250 iterations and a batch size of two. Finally, data augmentation techniques, including random rotation and scaling were added to enhance the robustness of training. All models in this study were trained from scratch with all parameters made trainable (i.e., no parameters were frozen), ensuring a fair and consistent evaluation of their performance on our specific task.

### 3.2.3 Ablation studies

In addition to comparing the six established U-shaped models, we also conducted ablation studies on them to further probe the relevant design choices that can influence their performance. First, to confirm the impact of attention mechanisms, we evaluated the variants with these DL models, all with four resolution stages. Second, to understand the impact of the number of resolution stages, we compared these models and their variants that are one stage different from the original architectures. Note that, here we evaluated 3DSwinUnet with three and four resolution stages using a patch embedding size of two, as the four-staged version has a downsampling limitation on the input patch size at the last resolution stage. Finally, the more nuanced design elements for U-shaped models, including the operations for skip-connections (SC), downsampling (DS), and upsampling (US) are often overlooked. To further improve the adapted 3DSwinUnet, which is a full Swin Transformer model and to better understand the effects of these operations, we investigated the influences of these design choices on the baseline 3DSwinUnet. Following the original 2D design, the baseline 3DSwinUnet employs linear layers in its skip-connection, downsampling, and upsampling. Here, we introduced four additional 3DSwinUnet variants, named 3DSwinUnetV1 $\sim$ 3DSwinUnetV4, with their detailed design differences listed in Table 3.4. For skip-connection and downsampling, we compared the application of linear and residual operations while for upsampling, we tested linear operation, nearest interpolation, and transpose convolution. Also, to investigate the impact of the bottleneck design in the 3DSwinUnet model, we have further introduced four variants based on 3DSwinUnetV4, including 3DSwinUnetB0 $\sim$ 3DSwinUnetB3. These variants explore the influence of employing no operation, one Swin Transformer block, and one or two Residual blocks at the bottleneck, as detailed in Table 3.5.


A
B
A∩B
DSC(A,B) =
2
+
= 2 |A∩B| / |A| + |B|
Maximum tolerated distance τ
A
Boundary of A, S_A
Border regions of A, ℬ_A^(τ)
Boundary outside of (S_A ∩ ℬ_B^(τ))
B
Boundary of B, S_B
Border regions of B, ℬ_B^(τ)
Boundary outside of (S_B ∩ ℬ_A^(τ))
Normalized Surface Distance (NSD)
NSD(A,B)^(τ) =
= |S_A ∩ ℬ_B^(τ)| + |S_B ∩ ℬ_A^(τ)| / |S_A| + |S_B|
Synonyms: Surface Dice, Normalized Surface Distance

(a) Dice Coefficient [69]

(b) Normalized Surface Distance (NSD) [69]

Figure 3.2: The formula of the segmentation accuracy metrics (Dice Coefficient and Normalized Surface Distance) used in this study.

### 3.2.4 Evaluation Metrics and Statistical Analysis

For each CT scan of the test cases, sliding windows with 50% overlapping were used to compute the automatic segmentation results, which were combined into one volume using a Gaussian weighting function with a standard deviation coefficient of 0.125. We compared the aforementioned DL models for their segmentation accuracy and efficiency for the Surgical Labels and BTCV Labels, which include anatomical structures of various sizes and geometries. Specifically, in terms of segmentation accuracy, we used the Dice coefficient (Fig. 3.2a) and Normalized Surface Distance (NSD) (Fig. 3.2b), both with the range of [0,1] (1 being the most desirable). While the first metric was widely used, it tends to favor larger objects and those with a bigger surface-to-volume ratio. To complement the Dice coefficient, the NSD measures the tightness of fit for the surfaces between the automatic segmentation and the ground truths. It can more appropriately assess the segmentation quality for smaller anatomies. Here, we used a $3mm$ threshold for computing the NSD.

In terms of computational complexity, we evaluated the number of parameters in the model and the inference latency. Specifically, the inference latency is the amount of time needed for a model to process one $96 \times 96 \times 96$ voxel image patch and was obtained by using the benchmark framework with 1000 run time provided by [34] on a desktop computer with an NVIDIA GeForce RTX 3090

GPU and an 11th Gen Intel® Core™ i9 CPU. The number of parameters reflects a model's complexity, correlating with its memory demand. At the same time, inference latency indicates the model's speed and computational demand. Considering both metrics helps us comprehensively assess the model's overall resource consumption, containing both memory and computational aspects.

To verify the differences in segmentation performances across different models, we performed a statistical analysis on the results. Specifically, two-way ANOVA tests were conducted to investigate whether group-wise difference exists among the tested models, and whether there exists a difference in terms of segmentation performance between the Surgical labels and BTCV labels. If the ANOVA test indicates a significant difference, pair-wise post-hoc analysis with Tukey's Honestly Significant Difference (HSD) was employed to further identify the between-model differences with statistical significance. Here, we defined a p-value of 0.05 as the threshold that indicates a statistical significance.

### 3.2.5 Algorithm Ranking Method

To properly assess the performance of algorithms [57], public medical image processing challenges have widely adopted various ranking methods that incorporate multiple evaluation metrics. In general, two main ranking mechanisms have been utilized in algorithm ranking:

- *Metric-based aggregation ("aggregate then rank")*: the evaluation metrics are first aggregated using median or mean across all cases, and then the ranking is based on the aggregated value.

- *Case-based aggregation ("rank then aggregate"*: the models are first ranked for each metric on each case, and then the average values of all the rankings are used to rank the models.

Maier *et al.* [57] found that metric-based aggregation is the most commonly used ranking method in different medical image processing challenges, and recommend using it over case-based aggregation. Furthermore, for metric-based aggregation, the mean as an aggregation method was evaluated to be more robust than the median. Following these guidelines, we adopted the "aggregate then rank" approach with the mean value for metric aggregation. Specifically, we first aggregated the results of all test cases and then ranked the models for each metric. With this strategy, we obtained the algorithm rankings for segmentation accuracy and model complexity by averaging the rankings

Table 3.1: Computational complexity and segmentation performance (mean±std) across 3DUNet, STUNet, AttentionUNet, and 3DSwinUnet models with different numbers of resolution stages.

(a) Computational Complexity and Accuracy (mean ± std)

| **Model Name** | **Model Complexity ↓** | | **Dice ↑** | | | **NSD ↑** | | |
|---|---|---|---|---|---|---|---|---|
| | **Parameters (M)** | **Inference Latency (ms)** | **BTCV** | **Surgery** | **Total** | **BTCV** | **Surgery** | **Total** |
| *3DUNet* | $30.6^{4}$ | $6.158^{1}$ | $93.66 \pm 3.08^{3}$ | $96.96 \pm 1.11^{2}$ | $95.18 \pm 1.69^{2}$ | $97.50 \pm 2.89^{2}$ | $98.88 \pm 1.25^{1}$ | $98.13 \pm 1.65^{2}$ |
| *STUNet* | $30.23^{2}$ | $7.298^{3}$ | $94.08 \pm 2.92^{1}$ | $97.04 \pm 1.20^{1}$ | $95.44 \pm 1.62^{1}$ | $97.57 \pm 2.54^{1}$ | $98.85 \pm 1.59^{2}$ | $98.16 \pm 1.54^{1}$ |
| *AttentionUNet* | $30.59^{3}$ | $7.193^{2}$ | $93.78 \pm 3.11^{2}$ | $96.76 \pm 1.29^{3}$ | $95.14 \pm 1.74^{3}$ | $97.40 \pm 2.84^{3}$ | $98.77 \pm 1.60^{3}$ | $98.03 \pm 1.69^{3}$ |
| *SwinUNETR* | $62.19^{6}$ | $18.585^{5}$ | $93.32 \pm 3.21^{5}$ | $96.54 \pm 1.45^{5}$ | $94.79 \pm 1.83^{5}$ | $97.10 \pm 2.86^{5}$ | $98.28 \pm 1.81^{5}$ | $97.64 \pm 1.78^{5}$ |
| *FocalSegNet* | $69.65^{7}$ | $15.412^{4}$ | $93.47 \pm 2.98^{4}$ | $96.57 \pm 1.39^{4}$ | $94.89 \pm 1.67^{4}$ | $97.20 \pm 2.81^{4}$ | $98.43 \pm 1.58^{4}$ | $97.77 \pm 1.66^{4}$ |
| *3DSwinUnet* | $7.98^{1}$ | $22.91^{6}$ | $46.15 \pm 7.02^{7}$ | $59.30 \pm 4.60^{7}$ | $52.20 \pm 5.06^{7}$ | $34.81 \pm 5.64^{7}$ | $38.52 \pm 4.15^{7}$ | $36.54 \pm 4.24^{7}$ |
| *3DSwinUnetV4* | $31.55^{5}$ | $29.025^{7}$ | $92.04 \pm 3.58^{6}$ | $95.72 \pm 1.58^{6}$ | $93.73 \pm 2.00^{6}$ | $96.14 \pm 3.47^{6}$ | $97.50 \pm 1.96^{6}$ | $96.76 \pm 2.06^{6}$ |

(b) Model Rankings

| **Model Name** | **Rankings ↓** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Model Complexity** | **Segmentation** | | | **Final** | | |
| | | BTCV | Surgery | Total | BTCV | Surgery | Total |
| *3DUNet* | 1 | 2 | 1 | 2 | 2 | 1 | 2 |
| *STUNet* | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| *AttentionUNet* | 1 | 2 | 2 | 3 | 2 | 2 | 3 |
| *SwinUNETR* | 3 | 4 | 4 | 5 | 4 | 4 | 5 |
| *FocalSegNet* | 3 | 3 | 3 | 4 | 3 | 3 | 4 |
| *3DSwinUnet* | 2 | 6 | 6 | 7 | 4 | 4 | 5 |
| *3DSwinUnetV4* | 4 | 5 | 5 | 6 | 5 | 5 | 6 |

of the sub-measures of each category. Finally, to fully consider both categories of factors, the final algorithm ranking was achieved based on the average of the segmentation accuracy ranking and model complexity ranking for each algorithm. For the assessment, we obtained the segmentation accuracy rankings and the final rankings for the Surgical label, BTCV labels, and the combination of both groups.

## 3.3 RESULTS

In Table 3.1, we summarize the segmentation accuracy, model complexity, and ranking for the 3DUNet, STUNet, AttentionUNet, SwinUNETR, FocalSegNet, and 3DSwinUnet. Additionally, the evaluation metrics of the ablation studies are listed in Tables 3.2 to 3.5. Finally, due to the page limit, we demonstrate the segmentation outcomes of these models with image examples and anatomy-wise boxplots in Fig. S3-S4 and S5-S6 of the *Supplementary Materials*, respectively.

### 3.3.1 Segmentation Accuracy

With the segmentation performance of the six U-shaped models, and the best-performing 3DSwinUnet (i.e., 3DSwinUnetV4) in Table 3.1, we found three general observations. First, these included models achieved significantly higher Dice scores on the surgical labels rather than BTCV ones ($p<0.05$). Second, the 3DSwinUnet model performed significantly worse in segmentation accuracy than the rest of the counterparts ($p<0.05$). Finally, despite the differences in model architectures, especially in the adoption of diverse attention mechanisms, and the variations in the mean metrics, their segmentation performances do not differ significantly ($p>0.05$). Specifically, for the surgical labels, the 3DUNet achieved the second best mean Dice score of 96.96% and the best mean NSD of 98.88% while the STUNet had the highest mean Dice score of 97.04% and the second best NSD of 98.85%, making them share the first ranking for segmentation accuracy in this task. The AttentionUNet, FocalSegNet, SwinUNETR, 3DSwinUnetV4, and 3DSwinUnet followed in our ranking, respectively. For the BTCV labels, the STUNet model also achieved the best Dice scores (94.08%) and NSD (97.57%). 3DUNet and AttentionUNet shared the second place in the ranking, with 3DUNet achieving the second best mean NSD (97.50%) while AttentionUNet had the second best Dice score of (93.78%). After them, the models in descending order of ranking are the FocalSegNet, SwinUNETR, 3DSwinUnetV4, and 3DSwinUnet. Finally, when pulling both sets of labels together, the STUNet model again achieved the highest Dice score of 95.44% and the best NSD of 98.16%, with the 3DUNet ranked second, followed by the AttentionUNet, FocalSegNet, SwinUNETR, 3DSwinUnetV4, and 3DSwinUnet.

### 3.3.2 Model Complexity

Primarily due to the choice of architecture types, the model complexity varies, with the 3DUNet having the lowest inference latency. By sharing the basic structure of the UNETR [31], the SwinUNETR and FocalSegNet have more than twice the computational cost of the CNN U-shaped models, with the latter containing the highest number of model parameters. Finally, 3DSwinUnet had the lightest model architecture, and 3DSwinUnetV4 had the highest inference latency on average.

### 3.3.3 Final algorithm ranking

In our final algorithm ranking for the total anatomical labels, the STUNet, 3DUNet, and AttentionUNet were ranked first, second, and third, respectively. These three models stood out due to their model efficiency and higher segmentation accuracy. Even though 3DUNet ranked second overall, it shared the first place with STUNet for surgical labels and second place with AttentionUNet for BTCV labels. The FocalSegNet, SwinUNETR, 3DSwinUnet, and 3DSwinUnetV4 were ranked next from the fourth to the last position for the combined total anatomical labels. As our ranking balances both accuracy and model complexity, although the baseline 3DSwinUnet had significantly lower segmentation performance ($P < 0.05$) than the rest, its lowest number of parameters boosted its final ranking.

### 3.3.4 Ablation studies

#### Impact of attention mechanisms

Table 3.2 provides the segmentation accuracy, model complexity, and ranking of the selected U-shaped models with four resolution stages. While our general observations align with those in Section 3.3.1, the model ranking changed slightly. Among the selected models, 3DUNet ranked first in all of our ranking categories thanks to its computational efficiency and high segmentation accuracy. Following 3DUNet, STUNet came as second in computation complexity, segmentation performance, and final ranking for the surgical labels. Meanwhile, the AttentionUNet ranked as the second best model in our segmentation and final rankings for the BTCV labels and combined labels. Together, FocalSegNet, SwinUNETR, and 3DSwinUnet placed as fourth in the computational complexity category. With the same number of resolution stages, 3DSwinUnet still kept the lowest number of parameters, and FocalSegNet had the lowest inference latency. Finally, FocalSegNet, SwinUNETR, and 3DSwinUnet placed from the fourth to the sixth in segmentation quality and final ranking for all three label groups, respectively.

Table 3.2: Model Performance (mean±std) and ranking across the selected four-stage U-shaped Models. Note that an individual metric's rankings are shown as superscripts beside the corresponding metrics.

(a) Computational Complexity and Accuracy (mean ± std)

| **Model Name** | **Model Complexity ↓** | | **Dice ↑** | | | **NSD ↑** | | |
|---|---|---|---|---|---|---|---|---|
| | **Parameters (M)** | **Inference Latency** (ms) | **BTCV** | **Surgery** | **Total** | **BTCV** | **Surgery** | **Total** |
| *3DUNet* | $14.59^{2}$ | $5.108 \pm 0.41^{1}$ | $94.03 \pm 3.02^{1}$ | $97.04 \pm 1.08^{1}$ | $95.41 \pm 1.65^{1}$ | $97.55 \pm 2.58^{1}$ | $98.90 \pm 1.36^{1}$ | $98.17 \pm 1.50^{1}$ |
| *STUNet* | $14.51^{1}$ | $6.052 \pm 0.65^{3}$ | $93.75 \pm 3.22^{3}$ | $96.87 \pm 1.14^{2}$ | $95.18 \pm 1.78^{3}$ | $97.41 \pm 2.78^{3}$ | $98.76 \pm 1.37^{2}$ | $98.03 \pm 1.59^{3}$ |
| *AttentionUNet* | $14.77^{3}$ | $5.970 \pm 0.44^{2}$ | $93.86 \pm 2.96^{2}$ | $96.82 \pm 1.38^{3}$ | $95.21 \pm 1.67^{2}$ | $97.48 \pm 2.77^{2}$ | $98.72 \pm 1.64^{3}$ | $98.05 \pm 1.70^{2}$ |
| *SwinUNETR* | $62.19^{5}$ | $18.585 \pm 1.24^{5}$ | $93.32 \pm 3.21^{5}$ | $96.54 \pm 1.45^{5}$ | $94.79 \pm 1.83^{5}$ | $97.10 \pm 2.86^{5}$ | $98.28 \pm 1.81^{5}$ | $97.64 \pm 1.78^{5}$ |
| *FocalSegNet* | $69.65^{6}$ | $15.412 \pm 0.84^{4}$ | $93.47 \pm 2.98^{4}$ | $96.57 \pm 1.39^{4}$ | $94.89 \pm 1.67^{4}$ | $97.20 \pm 2.81^{4}$ | $98.43 \pm 1.58^{4}$ | $97.77 \pm 1.66^{4}$ |
| *3DSwinUnet* | $30.91^{4}$ | $28.611 \pm 1.32^{6}$ | $59.71 \pm 6.12^{6}$ | $68.95 \pm 4.56^{6}$ | $63.98 \pm 4.35^{6}$ | $53.96 \pm 6.56^{6}$ | $58.23 \pm 4.59^{6}$ | $55.96 \pm 4.98^{6}$ |

(b) Model Rankings

| **Model Name** | **Rankings ↓** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Model Complexity** | **Segmentation** | | | **Final** | | |
| | | BTCV | Surgery | Total | BTCV | Surgery | Total |
| *3DUNet* | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| *STUNet* | 2 | 3 | 2 | 3 | 3 | 2 | 3 |
| *AttentionUNet* | 3 | 2 | 3 | 2 | 2 | 3 | 2 |
| *SwinUNETR* | 4 | 5 | 5 | 5 | 5 | 5 | 5 |
| *FocalSegNet* | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| *3DSwinUnet* | 4 | 6 | 6 | 6 | 6 | 6 | 6 |

### Impact of resolution stages

In Table 3.3, we show the segmentation accuracy results for 3DUNet, STUNet, AttentionUNet, and 3DSwinUnet across different numbers of resolution stages. The findings revealed distinct behaviors among the models when increasing the number of resolution stages. Specifically, increasing the number of resolution stages enhanced the performance of STUNet while it reduced the performance of 3DUNet and 3DSwinUnet. In the case of AttentionUNet, an increase in resolution stages led to a slight enhancement in model performance concerning the NSD score for surgery labels, while its performance was lower in all other scenarios. It is worth noting that despite the minor changes in the model performance due to varying the number of resolution stages, these changes were not statistically significant ($p > 0.05$).

### 3DSwinUnet modifications

Tables 3.4-3.5 summarize the computational complexity and segmentation accuracy for the corresponding operations used in each design component of the 3DSwinUnet variants. The findings

Table 3.3: Computational complexity and segmentation performance (mean±std) across 3DUNet, STUNet, AttentionUNet, and 3DSwinUnet models with different numbers of resolution stages.

| Model Name | Num of stages | Model Complexity ↓ | | Dice ↑ | | | NSD ↑ | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Parameters (M) | Inference Latency (ms) | BTCV | Surgery | Total | BTCV | Surgery | Total |
| *3DUNet* | 4 | 14.59 | 5.108 ± 0.41 | 94.03 ± 3.02 | 97.04 ± 1.08 | 95.41 ± 1.65 | 97.55 ± 2.58 | 98.90 ± 1.36 | 98.17 ± 1.50 |
| *3DUNet* | 5 | 30.6 | 6.158 ± 0.49 | 93.66 ± 3.08 | 96.96 ± 1.11 | 95.18 ± 1.69 | 97.50 ± 2.89 | 98.88 ± 1.25 | 98.13 ± 1.65 |
| *STUNet* | 4 | 14.51 | 6.052 ± 0.65 | 93.75 ± 3.22 | 96.87 ± 1.14 | 95.18 ± 1.78 | 97.41 ± 2.78 | 98.76 ± 1.37 | 98.03 ± 1.59 |
| *STUNet* | 5 | 30.23 | 7.298 ± 0.55 | 94.08 ± 2.92 | 97.04 ± 1.20 | 95.44 ± 1.62 | 97.57 ± 2.54 | 98.85 ± 1.59 | 98.16 ± 1.54 |
| *AttentionUNet* | 4 | 14.77 | 5.970 ± 0.44 | 93.86 ± 2.96 | 96.82 ± 1.38 | 95.21 ± 1.67 | 97.48 ± 2.77 | 98.72 ± 1.64 | 98.05 ± 1.70 |
| *AttentionUNet* | 5 | 30.59 | 7.193 ± 0.48 | 93.78 ± 3.11 | 96.76 ± 1.29 | 95.14 ± 1.74 | 97.40 ± 2.84 | 98.77 ± 1.60 | 98.03 ± 1.69 |
| *3DSwinUnet* | 3 | 7.85 | 22.860 ± 0.82 | 60.81 ± 6.24 | 70.76 ± 4.10 | 65.38 ± 4.26 | 54.48 ± 6.74 | 60.57 ± 4.70 | 57.29 ± 5.12 |
| *3DSwinUnet* | 4 | 30.91 | 28.611 ± 1.32 | 59.71 ± 6.12 | 68.95 ± 4.56 | 63.98 ± 4.35 | 53.96 ± 6.56 | 58.23 ± 4.59 | 55.96 ± 4.98 |

presented in Table 3.4 indicate that replacing linear with convolutional layers can improve model performance. Specifically, replacing the linear layers with residual blocks in skip-connection blocks in 3DSwinUnetV1 led to a significant ($p < 0.05$) performance improvement compared to 3DSwinUnet. For 3DSwinUnetV2 and 3DSwinUnetV3, replacing linear layers in decoder up-sampling blocks with nearest interpolation and transpose convolutional layers, respectively, resulted in a significant ($p < 0.05$) performance enhancement compared to 3DSwinUnetV1. However, there was only a slight difference ($p > 0.05$) between the performance of 3DSwinUnetV2 and 3DSwinUnetV3. Finally, using residual layers in encoder downsampling blocks resulted in a slight ($p > 0.05$) performance improvement in 3DSwinUnetV4 than in 3DSwinUnetV3.

Table 3.4: Computational complexity and segmentation performance (mean±std) across various 3DSwinUnet model variants. Here, SC=skip-connection type, DS=downsampling operation, and US=upsampling operation.

| Model Name | Details | | | Model Complexity ↓ | | Dice ↑ | | | NSD ↑ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | SC | DS | US | Parameters (M) | Inference Latency (ms) | BTCV | Surgery | Total | BTCV | Surgery | Total |
| *3DSwinUnet* | Linear | Linear | Linear | 7.98 | 22.910 ± 0.93 | 46.15 ± 7.02 | 59.30 ± 4.60 | 52.20 ± 5.06 | 34.81 ± 5.64 | 38.52 ± 4.15 | 36.54 ± 4.24 |
| *3DSwinUnetV1* | Residual | Linear | Linear | 24.39 | 27.382 ± 0.95 | 86.17 ± 5.50 | 92.04 ± 2.60 | 88.87 ± 3.24 | 89.06 ± 6.20 | 91.29 ± 3.58 | 90.09 ± 3.95 |
| *3DSwinUnetV2* | Residual | Linear | Interpolation | 23.57 | 27.716 ± 0.89 | 91.47 ± 3.89 | 95.32 ± 1.81 | 93.23 ± 2.22 | 95.76 ± 3.57 | 96.70 ± 2.32 | 96.19 ± 2.20 |
| *3DSwinUnetV3* | Residual | Linear | Transpose Convolution | 24.39 | 27.706 ± 1.42 | 91.67 ± 3.40 | 95.32 ± 1.62 | 93.34 ± 1.91 | 95.79 ± 3.36 | 96.65 ± 2.40 | 96.19 ± 2.11 |
| *3DSwinUnetV4* | Residual | Residual | Transpose Convolution | 31.55 | 29.025 ± 0.81 | 92.04 ± 3.58 | 95.72 ± 1.58 | 93.73 ± 2.00 | 96.14 ± 3.47 | 97.50 ± 1.96 | 96.76 ± 2.06 |

Further analysis of bottleneck modifications in Table 3.5 revealed a slight impact of different operations on 3DSwinUnetV4 performance ($p > 0.05$). Specifically, employing either a single Swin Transformer block or a Residual block in the bottleneck resulted in a marginal improvement

($p > 0.05$) in overall model performance (Total Dice and NSD). However, increasing the number of operations within the bottleneck, compared to using a single block, led to a minor decrease in the overall performance. The results revealed that using one Swin Transformer block in 3DSwinUnetB1 performed slightly better than its counterparts ($p > 0.05$).

Table 3.5: Comparison of Computational Complexity and Segmentation Performance Across 3DSwinunet Model Variants with Different Bottleneck Types.

| Model Name | Bottleneck Type | Model Complexity ↓ | | Dice ↑ | | | NSD ↑ | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Parameters (M) | Inference Latency (ms) | BTCV | Surgery | Total | BTCV | Surgery | Total |
| 3DSwinUnetB0 | None | 27.89 | $26.21 \pm 2.94$ | $91.87 \pm 3.70$ | $95.75 \pm 1.61$ | $93.65 \pm 2.04$ | $96.07 \pm 3.41$ | $97.50 \pm 2.13$ | $96.73 \pm 2.01$ |
| 3DSwinUnetB1 | Swin Transformer | 29.72 | $28.04 \pm 3.22$ | $91.99 \pm 3.70$ | $95.83 \pm 1.46$ | $93.75 \pm 2.03$ | $96.19 \pm 3.50$ | $97.69 \pm 1.84$ | $96.88 \pm 2.05$ |
| 3DSwinUnetV4 | Swin Transformer×2 | 31.55 | $29.025 \pm 0.81$ | $92.04 \pm 3.58$ | $95.72 \pm 1.58$ | $93.73 \pm 2.00$ | $96.14 \pm 3.47$ | $97.50 \pm 1.96$ | $96.76 \pm 2.06$ |
| 3DSwinUnetB2 | Residual | 35.85 | $26.96 \pm 3.11$ | $92.06 \pm 3.66$ | $95.74 \pm 1.69$ | $93.75 \pm 2.06$ | $96.14 \pm 3.88$ | $97.53 \pm 2.17$ | $96.77 \pm 2.33$ |
| 3DSwinUnetB3 | Residual×2 | 43.82 | $27.98 \pm 3.08$ | $91.92 \pm 3.76$ | $95.83 \pm 1.48$ | $93.71 \pm 2.10$ | $96.07 \pm 3.48$ | $97.65 \pm 1.82$ | $96.79 \pm 2.03$ |

## 3.4 Discussion

We compared six state-of-the-art U-shaped techniques and one model variant (3DSwinUnetV4) with a focus on the impacts of different attention mechanisms, including attention gates, self-attention, and focal modulation, as well as comparing CNN-Transformer hybrid and full Transformer models. We didn't observe a significant difference ($p>0.05$) among these models, except for the 3DSwinUnet. However, the attention mechanisms were ranked in the descending order of attention gate, focal modulation, window-based self-attention, and full integration of window-based self-attention. While some previous studies showed the benefit of various attention mechanisms in U-shaped models [81], in some recent medical image segmentation benchmarking reports [44, 25], 3D UNets were shown to perform better than CNN-Transformer hybrid counterparts. To further confirm this observation while removing the influence of different resolution stages [46], we conducted an ablation study by fixing the number of resolution stages of all models to four, and the observation remained. This could potentially be explained by the patch-based training and inference, where limited field-of-view and information redundancy may benefit local feature extraction slightly. The segmentation accuracy ranking of the attention mechanisms reflected their ability to encode local features. Furthermore, the observation may also be due to the need for larger training datasets for Transformer models. In clinical deployment, computational efficiency is important, so with similar segmentation accuracy, CNN U-shaped models that contain fewer parameters and faster inference latency can be more favorable. Across all tested models in Table 3.1, we demonstrated that the segmentation quality is in general better for larger anatomies. This is reflected in the significant differences ($p<0.05$) of Dice coefficient and NSD between the BTCV and surgical labels, where the latter contains larger anatomical structures. This observation is consistent with previous studies [44]. Furthermore, for these smaller anatomies, the standard deviations of the segmentation accuracy metrics are also greater than those of the larger structures, suggesting lower robustness.

Deeper layers in convolutional neural networks may help encode more refined features for relevant tasks [80]. However, based on Table 3.3, more resolution stages in U-shaped models do not always improve segmentation accuracy. Among the models, STUNet was the only model that slightly benefited from increasing the number of resolution stages potentially because of its extensive incorporation of residual blocks and multi-scale processing in its architecture design. For pure Transformer-based models, increasing the number of resolution stages may reduce performance due to the loss of spatial information in the last layers and cause the attention to collapse [95]. To mitigate this, the integration of the attention mechanism with convolutional operations, as proposed in paper [18] can be beneficial. Finally, as additional resolution stages augment model complexity, the advantage of large models could also depend on the complexity of the task, size of training data, and the initial input image size, and resolution.

The optimization of model complexity revealed a trade-off between the number of parameters and inference latency, with an interplay with other additional factors. First, incorporating interpolation operations for upsampling in both CNN-based and attention-based U-shaped models, while reducing the number of parameters, resulted in increased inference latency. This was evident in the STUNet model, which, despite having fewer parameters than 3DUNet, showed higher latency. Similarly, within the attention-based U-shaped models, 3DSwinUnetV2 demonstrated higher latency despite having fewer parameters than its counterparts, 3DSwinUnetV1 and 3DSwinUnetV3, due to the inclusion of interpolation layers. Second, within the attention-based U-shaped models, incorporating self-attention mechanisms resulted in longer inference latencies than focal modulation (i.e., FocalSegNet), even though the self-attention-based models had fewer parameters. This suggests that the computational demands of the self-attention mechanism, such as computing similarity coefficients and employing a window-based mechanism, contribute to the increased latency. In contrast, the FocalSegNet's use of depthwise convolutions, while increasing the parameter count, led to reduced inference latency compared to self-attention mechanisms. This highlights the impact of specific computational operation choices on model complexity attributes.

Finally, a direct adaptation of the original 2D SwinUnet model [9] for 3D segmentation offered sub-optimal performance as previously mentioned in another study [32]. Although the shifted window mechanism in the Swin Transformer adds inductive bias to this operation, its self-attention can still interfere with local information, as previously noted in [33]. The lack of local feature extractors in the 3DSwinUnet, due to the use of Swin Transformer and linear layers, results in suboptimal performance in our analysis. Previous studies [63, 32] attempted to address this issue by incorporating the cross-attention mechanism into the decoders' Swin Transformer blocks. Although QTUNet [32] showed improved performance over VTUNet [63] in the BTCV dataset, the results on this dataset remained suboptimal and required further investigation. Therefore, in our work, we diverge from the Cross-Attention approach and explore using convolutional operations to enhance local feature representation in SwinUnet for 3D image analysis of thoracic organs. Inspired by the use of residual blocks in the STUNet, we developed four variants of the 3DSwinUNet to fully explore the potential of pure Transformer UNets by modifying the operations of upsampling, skip-connection, and downsampling (Table 3.4).

In short, we found greatly enhanced performance when replacing linear operations with residual blocks in skip connections and downsampling operations, as well as employing interpolation or transpose convolution in upsampling operations for the baseline 3DSwinUnet. As residual blocks are known to improve the learning efficiency of hierarchical features and gradient flow during training, they could potentially promote inductive bias compared with simple linear operations. Notably, the performance boost in terms of accuracy and robustness (i.e., lower standard deviations of metrics) was much more evident for their deployment in skip connections than downsampling. While keeping residual blocks for skip-connections and linear operation for downsampling (3DSwinUnetV2 vs. 3DSwinUnetV3), upsampling operations with interpolation and transpose convolution provided similar further accuracy enhancement, with the latter offering slightly better performance robustness. This enhancement was due to the hierarchical representation of convolutional layers [96].

Given that prior studies that modified SwinUnet for 3D image analysis often incorporated a Swin Transformer block at the bottleneck of their models [63, 32] we investigated the impact of different bottleneck configurations on our 3DSwinUnetV4 (see Table 3.5), the top-performing variant from our experiments. Our results demonstrate that using either a Swin Transformer or Residual block at the bottleneck is effective, likely due to their ability to extract fine-grained features from the down-sampled encoder output. Notably, the Swin Transformer block consistently outperformed the Residual block. This may be attributed to the fact that in 3DSwinUnetV4, the bottleneck receives input features already processed by residual blocks within the patch merging blocks. Consequently, employing a Swin Transformer block after the residual blocks can enhance results by leveraging its self-attention mechanism to extract global features. This is consistent with findings in [62, 33], where self-attention after convolutional operations aggregated features and improved robustness to high-frequency noise, similar to spatial smoothing. Further experiments increasing the number of blocks in the bottleneck yielded mixed results in different datasets. Increasing Swin Transformer blocks in 3DSwinUnetV4 led to a higher BTCV Dice score compared to 3DSwinUnetB1 (which has one Swin Transformer block). However, overall performance decreased, potentially due to the loss of spatial information and attention collapse in the deeper layers [95]. On the other hand, Increasing the number of Residual blocks in 3DSwinUnetB3 also resulted in mixed outcomes, with small improvements in the surgical labels but a minor decrease in performance on BTCV. This suggests the potential benefit of deeper residual blocks, but the optimal structure may be task-dependent. Finally, our findings suggest that while increasing the depth of the bottleneck could potentially enhance results, this effect was inconsistent across tasks and led to increased model complexity. Therefore, we did not find increasing the depth of the model bottleneck to be efficient, considering the trade-off between accuracy and complexity. Therefore, we recommend using a single block at the bottleneck of the model, which extracts different features than the final encoder layer (global feature extractor if the previous layer is a local feature extractor, or vice-versa)

Our presented study primarily focuses on the impacts of attention mechanisms, number of resolution stages, and network configuration designs for U-shaped deep learning models. As a result, we selected the most popular and representative models for the themes of our investigation, instead of analyzing an exhaustive list of U-shaped models. For the model performance, it is possible that the insights drawn from the experiments may be task- and modality-specific [90, 51], and should be verified further with additional benchmarking datasets.

## 3.5 Conclusion

In this study, we conducted a comprehensive evaluation of various U-shaped deep learning models in CT-based segmentation for thoracic surgical planning and other abdominal anatomies (BTCV dataset), and showed that the STUNet ranked the best for the designated tasks based on the joint consideration of accuracy and model complexity. In summary, we found that CNN U-shaped models offer excellent values for the demonstrated tasks while attention mechanisms may not necessarily enhance the outcomes, with those better preserving local features gaining a slight edge in patch-based processing. In addition, although augmenting resolution stages does not always result in better accuracy, careful design of operations for different components of the U-shaped models can greatly boost the results. We hope the insights from our experiments will facilitate the deployment and development of deep learning models for the demonstrated application and beyond. In our future work, we aim to further expand this study by evaluating the models on more diverse benchmarking datasets and exploring the effects of various skip-connection mechanisms, such as nested hierarchical and Transformer-based approaches. Also, we will investigate the optimal integration of Transformer blocks within CNN models to enhance segmentation accuracy, as well as explore the impact of incorporating different attention blocks, like squeeze-and-excitation [37] or MRA block [20], within U-shaped architectures.

# Chapter 4

# Virtual Reality-Based Preoperative Planning for Optimized Trocar Placement in Thoracic Surgery: A Preliminary Study

A version of this chapter has been presented in Augmented Environments for Computer Assisted Interventions (AE-CAI 2024) joint workshop at the Medical Image Computing and Computer Assisted Interventions (MICCAI 2024) Conference. The conference proceeding will be published in the Wiley journal *Technology Healthcare Letters*:

- Harirpoush A, Rakovich G, Kersten-Oertel M, Xiao Y. Virtual Reality-Based Preoperative Planning for Optimized Trocar Placement in Thoracic Surgery: A Preliminary Study. *Technology Healthcare letters*, in press (arXiv preprint arXiv:2409.04414), 2024 [28].

## 4.1 Introduction

Lung cancer is the second most common cancer and the leading cause of cancer-related deaths worldwide [79]. In the United States, approximately 56,000 to 57,000 lung cancer resections are performed each year, with lobectomies being the most common type of resection [64]. Low post-trauma minimally invasive surgeries, such as video-assisted thoracoscopic surgery (VATS), are now being used to treat early-stage non-small-cell lung cancer [5]. During VATS surgeries, optimal trocar placement, which guides the entry of surgical tools and endoscopic camera into the body through small incisions is necessary for surgical success. Optimal placement involves three key principles: (1) Trocars must be carefully positioned to ensure full access to all relevant areas within the thoracic cavity to facilitate complete surgical exploration and intervention. (2) The endoscopic camera trocar should be strategically placed to provide a panoramic view of the surgical field and sufficient room for instrument manipulation and avoiding visual obstruction. (3) All trocar placements should be meticulously planned to prevent instrument crowding or "fencing", ensuring smooth and efficient instrument handling throughout the procedure [73, 49].

While VATS offers numerous benefits, the optimal placement of trocars remains an area of limited research and standardized guidelines. Two common principles to guide trocar placement exist: (1) the Baseball Diamond Principle (BDP), which offers enhanced maneuverability and wider access to the thoracic cavity, particularly advantageous in non-pulmonary procedures [42], and (2) the Triangle Target Principle (TTP), which optimizes direct access to the surgical target and is preferred for retraction or stapling [73, 42]. Despite these principles, surgeons primarily rely on their experience and patient-specific anatomy to make trocar placement decisions [73], potentially leading to longer operating times, increased risk of complications, and greater fatigue for the surgical team due to limited instrument working area, and maneuverability [65]. Thus there is a need for effective preoperative planning techniques, such as through virtual reality (VR) for precise and effective trocar placement.

In this paper, we introduce the first VR application for thoracic pre-operative planning to efficiently provide optimal trocar placement based on established surgical principles and developed in close collaboration with an experienced thoracic surgeon. In a preliminary study, we showcase the system's application in right upper lung lobectomy, a common thoracic surgery. Following conventional practice, we included three trocars: two for surgical instruments in tissue resection and manipulation and one for the insertion of an endoscopic camera for surgical monitoring. The importance of accessing all areas of the chest cavity in this procedure led to the development of a rule-based trocar placement system. This system aims to help in precise trocar placement to optimize the operable area, i.e., the intersection between the working area of surgical instruments and the endoscopic camera's field of view (FOV).

We designed three key VR interaction and visualization features that are tailored for thoracic surgery. **First**, to enhance precision in planning, our application uses a pivot mechanism for surgical tool trocar placement. **Second**, we employed a ”hand grabbing” interaction method for endoscopic camera position planning and camera trocar placement. **Lastly**, real-time visual feedback and evaluation metrics were devised to further assist in trocar placement based on existing guidelines and discussions with an experienced thoracic surgeon. Upon completion of planning, a comprehensive summary is generated, detailing key metrics for surgical plan quality to allow further refinement of plans. A preliminary user study was done to confirm the system's robustness and usability. The resulting insights can provide valuable information for future development of VR surgical applications for thoracic procedures and beyond.

## 4.2 Related Works

### 4.2.1 Patient-specific 3D Models

Recent studies have highlighted the significant advantages of incorporating patient-specific 3D models into preoperative planning across various surgical specialties [11, 2, 82, 65]. Within thoracic surgery, Cen et al. [11] demonstrated the utility of both physical (3D printed) and digital (VR/MR) 3D models in improving surgical field alignment during complex pulmonary atresia surgeries [11]. Ujiie et al. [82] focused on lung segmentectomy, utilizing a VR-based system with patient-specific 3D lung models to enhance surgical planning and surgeon confidence by facilitating the identification of anatomical landmarks and potential surgical challenges.

The value of 3D models extends beyond thoracic procedures. In laparoscopic hiatal hernia repair, Preda et al. [65] developed a preoperative planning system based on patient-specific 3D reconstruction and simulation, receiving positive feedback from surgeons who noted its potential to improve ergonomics and its particular value in challenging cases involving obese patients with large hiatal hernias. Further evidence for the utility of 3D models in thoracic surgery comes from Bakhuis et al. [2], who compared 2D planning with CT images to 3D planning in VR for pulmonary segmentectomy. Their findings revealed that 2D planns were adjusted in 52% of cases and tumor localization was inaccurate in 14%, underscoring the potential of 3D models to improve surgical accuracy and planning [2]. Beyond their use in individual procedures, Heuts et al. [36] explored the broader benefits of 3D models in thoracic surgical planning, finding that their use increases surgical efficiency, minimizes complications, and enhances overall surgical outcomes [36].

### 4.2.2 Extended Reality Applications in Minimally Invasive Surgeries

Extended Reality (XR) has been used in various minimally invasive surgeries to enhance procedural efficiency and precision. Several studies [54, 21, 22] have explored the use of XR for trocar planning systems to optimize minimally invasive surgery outcomes. For instance, López-Mir et al. [54] developed an augmented reality (AR) system to improve trocar placement accuracy in laparoscopic cholecystectomy, which is facilitated by a full HD monitor with transparency for enhanced depth perception. In their study involving four clinicians and 24 patients, the AR system demonstrated a 33% improvement in accuracy compared to traditional trocar placement methods. Similarly, Feuerstein et al. [21] presented an AR system for port placement in robotic-assisted surgeries (RATS). Their approach involved registering the patient for their preoperative CT scan by maneuvering the endoscope around fiducials, enabling automatic 3D position reconstruction. Later, Feuerstein et al. [22] proposed an AR system for port placement and intraoperative planning in minimally invasive liver resection that further accounts for intraoperative organ shifts. In another study, Bauernschmitt et al. [4] reported a significant reduction in operation time in minimally invasive robot-assisted heart surgery, thanks to employing their AR system for offline port placement planning and intraoperative navigation. Meanwhile, other endeavors [75, 74] have proposed decision-based mixed-reality (MR) and AR systems for automatic path planning to enhance surgical performance and streamline surgical workflows. For example, Simoes and Cao [75] introduced a decision-aid MR system to improve RATS performance and reduce planning time. Their system incorporates an optimization algorithm that suggests trocar placements based on the patient's anatomy and the specific surgery type. These suggestions are then projected onto the patient's body with a projector, allowing surgeons to refine the placement as needed. In another study, Schwenderling et al. [74] proposed a condition-based automated path planning AR system for percutaneous interventions. This system uses a projector to visualize the insertion point, path quality, and target on a phantom. Their results demonstrated the potential of visualizing insertion points and path quality in selecting safer access paths.

Beyond surgical planning, virtual Reality (VR) environments with haptic feedback devices have emerged as valuable tools for simulating surgical procedures and training trocar placement. Addressing limitations in previous training modules, such as limited anatomical variation, Solomon et al. [77] proposed a VR training system with haptic feedback to simulate VATS right upper lobectomy. In their system, trocar placement for each instrument is selected from predetermined sites on the chest wall, and instruments are then controlled via haptic devices. The process begins with determining the 30-degree thoracoscope trocar location, followed by an inspection of the anatomy through a camera view to guide the placement of the remaining trocars. The system includes both training and testing modes, with the latter featuring pop-up questions and explanations for incorrect answers. Similarly, Haidari et al. [26] developed a VR system with haptic devices for simulating VATS resection of the five lung lobes. Their study involved surgeons across three experience levels: novice, intermediate, and experienced. Their results showed significant differences between novices and experienced surgeons in blood loss, procedure time, and total instrument path length. Meanwhile, the only significant difference between intermediates and experienced surgeons was in procedure time.

While previous studies have widely investigated the influence of XR environments and patient-specific 3D models in surgical planning, the use of HMD VR systems for trocar placement in VATS remains untouched. This method could enhance surgical outcomes by offering surgeons superior depth perception and spatial understanding compared to traditional AR-based or monitor-based methods. Furthermore, using a VR environment could decrease potential registration errors that may arise in AR systems, thereby contributing to increased precision in surgical planning.

## 4.3 Materials and Methods

### 4.3.1 3D Model Generation

A 3D thoracic anatomical model was constructed based on a patient computed tomography (CT) scan (1.5×1.5×1.5 $mm^3$ resolution) selected from the publicly available TotalSegmentator [88] dataset. We obtained anatomical segmentations of the vertebrae, ribs, scapula, and trachea, which were manually refined using 3D Slicer to enhance model accuracy. Additionally, we further manually segmented the pulmonary vasculature and skin surface with 3D Slicer. All segmentations were converted into triangulated meshes (.obj format), and then integrated into the VR environment.

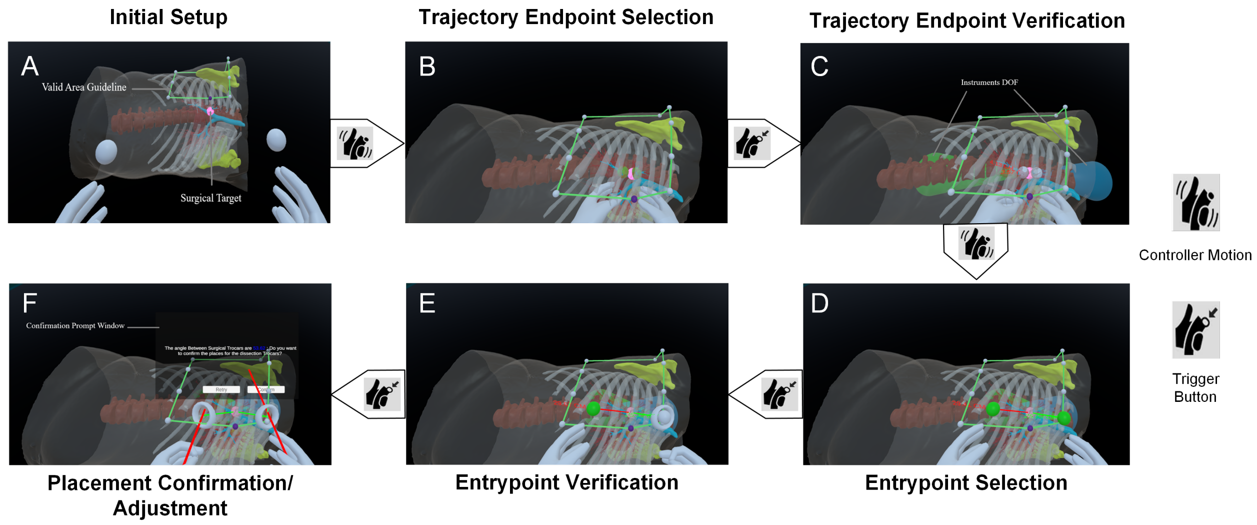

Initial Setup
Trajectory Endpoint Selection
Trajectory Endpoint Verification
A
Valid Area Guideline
Surgical Target
B
C
Instruments DOF
Controller Motion
Trigger Button
F
Confirmation Prompt Window
The angle Between Surgical Trocars are 53.62 ; Do you want to confirm the places for the dissection Trocars?
Retry
Confirm
E
D
Placement Confirmation/ Adjustment
Entrypoint Verification
Entrypoint Selection

Figure 4.1: Overview of the pivot mechanism in surgical trocar placement: A. Initial anterior view with trajectory endpoint spheres positioned in front of each controller; B. Spheres manipulated to define endpoints (green when near target); C. Endpoint verification displays working area and trajectory paths; D. Spheres moved to the skin to define entry points (green on contact); E. Green spheres and paths indicate valid entry, verifying trocar placement; F. Manipulation angle displayed for adjustment/confirmation.

### 4.3.2 VR user interface and workflow

Our system was created using the Oculus Quest Pro headset and controllers, employing the Unity game engine (Version 2021.3.11f1). Both development and user studies were conducted on a desktop computer with an NVIDIA GeForce RTX 3090 GPU, an 11th Gen Intel® Core™ i9 CPU, and 32 GB of RAM. The VR environment developed for this study includes three main visual components. **First**, a large information panel is positioned in front of the user to provide instructions for surgical planning tasks. **Second**, a virtual screen is positioned to the right of the information panel to display simulated video streaming from the virtual endoscopic camera, enabling precise adjustments and optimal positioning of the camera. **Third**, a detailed 3D anatomical model, featuring distinctly color-coded anatomical structures (see Fig. 1A, vertebrae in brown, scapula in yellow, trachea in blue, and pulmonary vasculature in red) is placed in front of the user for surgical planning. In the 3D model, we annotated the convergent point of the surgical tool trajectories and the optical axis of the endoscopic camera as a pink sphere. This convergent point was identified by our collaborating surgeon as the root of the right upper lobe and is common for planning most lung procedures. As key anatomies in surgical planning, we render the skin and ribs as semi-transparent structures to allow views of the underlying anatomy and their spatial relationship.

The workflow of the system is as follows. During the surgical planning, the user will remain in a standing position, mimicking a surgeon's posture during surgery. Before initiating planning, the user is asked to re-adjust the vertical position of the anatomical model to a comfortable level by using a slider selection tool shown in a control panel in the VR environment. Afterwards, planning can be initiated by pressing the "Start" button on the control panel. Typically during the right upper lung lobectomy procedure, the surgeon operates from the front of the patient (anterior view) while the camera-holding assistant is positioned at the back (posterior view). Therefore, the positioning of the patient model will be automatically adjusted according to this convention for the two surgical planning tasks in sequence: (1) surgical tool trocar placement with an anterior view of the patient, replicating the surgeon's perspective, and (2) endoscopic camera and the associated trocar placement, with a posterior view that mirrors the assistant's perspective. This task sequence was refined through an iterative development process to enhance workflow efficiency. In both tasks, the system provides visual feedback as color cues and numerical metric displacement in VR to guide users toward valid trocar placement areas. Further details on the data visualization and interaction schemes are provided in Sections 4.3.2 and 4.3.2.

### Surgical tool trocar placement

The trocar placement uses a pivot mechanism guided by two white spheres, one attached at the tip of each controller (Fig. 4.1A). This mechanism consists of two phases: *endpoint selection* and *entry point placement* for the surgical trajectories. **First**, the user reaches the two white spheres from left and right controllers within a 3D anatomical model towards the convergent point (the pink sphere) until the sphere turns green (Fig. 4.1B) indicating correct endpoint localization. The endpoints (i.e., white spheres) are placed by pressing the corresponding controller's trigger button. Afterward, a red surgical trajectory line will extend from the placed endpoint to each controller, along with a 20-degree-angle cone, the angle between the side to the principal axis, that represents the degree-of-free (DOF) of the surgical instrument's maneuver. The cone angle was defined using the surgeon's wrist range of motion (40 degrees for radial-ulnar deviation), as indicated by previous research [71]. Note that the right trocar's DOF cone is indicated by green color and the left one's by blue (Fig. 4.1C).

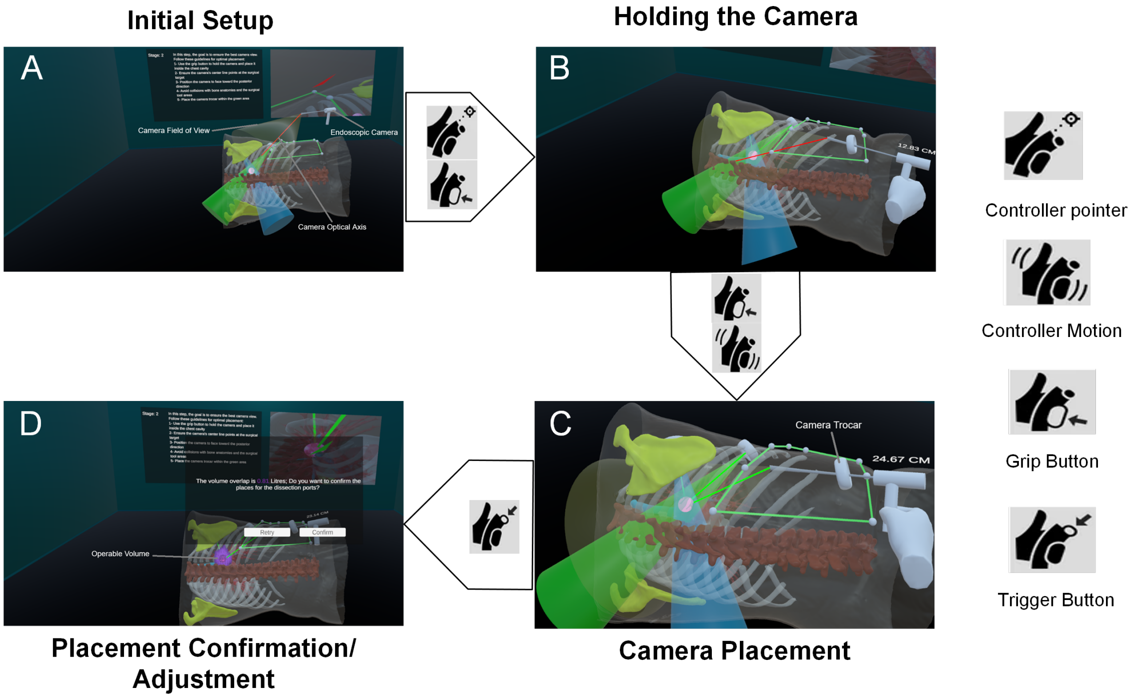
**Second**, the user drags the trajectory lines with the controllers onto the skin surface while ensuring that they avoid bony structures and that the real-time displayed trajectory distance for each controller remains under 28 cm, which is the maximum working length of the surgical instruments. The user must place the trocars in the designated area as contoured by green lines on the anatomical model. When these criteria are met, the system provides visual cues by turning both the trajectory lines and spheres green (Fig. 4.1D). The user then fixes the placement of each of the two trocars by pressing the corresponding controller's trigger button (Fig. 4.1E). After fixing both trocars, the "manipulation angle" between the two trajectories is displayed on a confirmation panel to confirm the planning or repeat the procedure till satisfaction (Fig. 4.1F). Note that prior research [27] suggests a manipulation angle between 45 and 75 degrees for optimal surgical instrument positioning with trocars parallel and sufficiently spaced.

### Endoscopic camera placement

For our system, we simulate a rigid endoscopic camera (an elongated tube with the camera at the tip) with a 30-degree tilt angle (between the optical axis and the rigid tubular body of the camera) and a 60-degree field of view, which is preferred for thoracic surgery [55]. During the task of endoscopic camera placement, we visualize the camera's FOV as a semi-transparent yellow cone and the optical axis as a red line (Fig. 4.2A). The user can manipulate the camera using a hand-grabbing interaction, by pressing the grip button of their dominant controller to hold and release it to place it in space(Fig. 4.2B). The second task of surgical planning requires the user to insert the camera into the chest cavity, by aiming the optical axis towards the convergent point (pink sphere) and checking the virtual camera display for optimal views. Upon inserting the camera tube into the body, a virtual trocar appears intersecting the skin surface, marking the camera's entry point and guiding the user to position it within the designated area (as contoured by green lines on the anatomical model). To avoid instrument crowding, the camera should be positioned outside the working area (shown as blue and green cones) of the surgical tools. Further, contact with bony structures should be avoided. To ensure correct placement, the red line (camera optical axis) will turn green once it aims directly at the convergent point without obstructions (Fig. 4.2C). Upon pressing the trigger button of the controller, a confirmation panel will appear to confirm or repeat the placement. Upon confirmation, the operable volume that considers the surgical tools' DOFs and camera's FOV will be calculated and visualized as purple voxels with numerical quantification in liters (Fig. 4.2D).

### 4.3.3 Computing operable volume

For the surgery, it is desirable to maximize the area that both surgical tools can cooperate while the endoscopic camera can inspect the full operation of the tools. Thus, the operable volume is determined by the overlap between the surgical tools' DOF and the camera's FOV, represented as three different cones. While triangulated meshes accurately represent the surface of objects, they do not provide the volume of the mesh. To address this, we employed the mesh voxelization method introduced by [23] to compute the operable volume. This consists of three steps: (1) A 3D grid surrounding the given mesh will be created, forming the foundation for the process with each cell representing a voxel. (2) The mesh surface will be voxelized by identifying voxels intersecting with the mesh triangles, effectively replacing the triangulated representation with small 3D cubes. (3) A scan-line fill algorithm will be used to identify the voxels within the object border. This process



Stage: 2 In this step, the goal is to ensure the best camera view. Follow these guidelines for optimal placement:
1- Use the grip button to hold the camera and place it inside the chest cavity
2- Ensure the camera's center line points at the surgical target
3- Position the camera to face toward the posterior direction
4- Avoid collisions with bone anatomies and the surgical tool areas
5- Place the camera trocar within the green area

Stage: 2 In this step, the goal is to ensure the best camera view. Follow these guidelines for optimal placement:
1- Use the grip button to hold the camera and place it inside the chest cavity
2- Ensure the camera's center line points at the surgical target
3- Position the camera to face toward the posterior direction
4- Avoid collisions with bone anatomies and the surgical tool areas
5- Place the camera trocar within the green area

The volume overlap is 0.81 Litres; Do you want to confirm the places for the dissection ports?

Figure 4.2: Overview of the hand grabbing method in camera placement: A. Initial posterior view and endoscopic camera; B. Pointing toward endoscopic camera and hold it by pressing grip button; C. Green camera optical axis line demonstrates valid placement; D. Volume of operable area displayed for adjustment/confirmation.

is similar to filling a shape in 2D by drawing horizontal lines until the boundaries are reached. To balance accuracy and efficiency, we use $1.5cm \times 1.5cm \times 1.5cm$ voxels; smaller voxels would improve resolution but increase computational cost. We customized the implementation of Mattatz [58], which was based on the work of [23] to compute the operable volume. Specifically, to compute the volumetric overlap between multiple meshes, we use a single 3D grid covering all models. Each mesh is assigned a unique ID (one for each cone), and for each voxel, the mesh ID is stored in a HashSet. Overlapping voxels are identified by HashSets containing the same number of elements as the input meshes.

### 4.3.4 User study design & system validation

Upon informed consent, we recruited 20 non-clinician participants (age = 25.95 ± 3.31 years, 7 female, 13 male) for our user study. To better understand the study cohort, we also surveyed their level of familiarity with VR technology and human anatomy. Among them, 75% indicated ”Familiar” or ”Somewhat Familiar” with VR, while only 30% reported similar familiarity with human anatomy, with one participant indicating ”Unfamiliar” with both. All participants were right-handed, and two (one male, one female) reported color blindness. No participants experienced VR sickness.

Participants were first given a brief Powerpoint presentation introducing the clinical context, tasks, and goals of the study. Following this, a hands-on tutorial was conducted to familiarize participants with the VR environment, planning process, and various interactions. This tutorial involved tasks different from those in the main study. During the tutorial, participants practiced planning on the left side of the 3D patient model, with an anterior view provided. Text-to-speech technology for the instruction from the information panel was integrated to offer assistance throughout each task. For the camera placement task, a semi-transparent ”phantom camera” positioned at the desirable location and position was presented as a ground truth reference, and the participants were asked to place the actual camera to overlap with the phantom guide. This served to illustrate optimal camera placement and angling towards the posterior side of the patient, as required in the surgery. Participants were encouraged to continue practicing until they felt comfortable using the system. Following the tutorial, we conducted the user study to formally validate our proposed system by following the workflow introduced in Section 4.3.2.

The proposed system was evaluated through a mixed-methods approach employing both semi-quantitative and quantitative measures. System usability was assessed using the System Usability Scale (SUS) by Brooke et al. [7], a widely recognized standardized questionnaire. The SUS evaluation is a Likert-scale questionnaire consisting of ten items, each with a range of 1 (strongly disagree) to 5 (strongly agree) [50]. Questions alternate between positively and negatively worded statements, ensuring participants actively engage with the content and thoughtfully consider their responses. These questions cover various aspects of the system, including effectiveness, efficiency, and overall user satisfaction. Among the 10 questions of SUS, each odd-numbered question is scored as x-1, and each even-numbered question is scored as 5-x, where x is the question’s resulting value. The scores for each participant are then summed, and then multiplied by 2.5 - resulting in a maximum SUS score of 100. A software system that receives an SUS score above 68 indicates good usability.

To further evaluate participant experience and effectiveness of the tailored data visualization and interaction designs, an additional Likert-scale questionnaire with eleven items was used to assess engagement, immersion, system usability, and the efficacy of visualizations, interactions, and visual feedback (the questions are detailed in Fig. 4.4). Specifically, the participants were asked to evaluate their engagement level within the application, the application's visual appeal, and usefulness in the designated task as well as the ergonomic design of the system. They were also asked to evaluate the ease of use and effectiveness of specific functionalities, including pivoting methods for surgical trocar placement, the hand-grabbing for camera placement, the visual feedback mechanisms provided, the information panels, and the final visualization of the operable volume. Participants rated each item on a 1-to-5 Likert scale (1=strongly disagree, 5=strongly agree). Finally, participants were asked to provide open-ended feedback on the positive and negative aspects of the system, along with recommendations for system improvement, and reported their familiarity with virtual reality (VR) and human anatomy. For the total SUS score, a one-sample t-test was used to assess whether the results were significantly different from 68. For each SUS sub-score and the customized UX questions, we compared the results to a neutral response (score=3), also with the Mann–Whitney U test. A $p - value < 0.05$ was used to indicate a statistically significant difference.

In addition to the semi-quantitative assessment, relevant quantitative metrics were collected from the proposed VR system for each designated task. These metrics included the total time spent on each task, the number of adjustments made in each task, and the historical and final positions of the trocars and the camera. For the first task (surgical trocar placement), we also recorded trajectory distance (in cm) for each surgical instrument (measured as the distance between the skin entry point and the surgical target), as well as the manipulation angle (the angle between the instruments upon reaching the surgical target). For the second task (camera placement), the volume of overlap between the camera's field of view and the surgical instruments' working area (in liters) was recorded.

## 4.4 Results

### 4.4.1 Semi-Quantitative Evaluation

Our VR system achieved an average SUS score of $81.8 \pm 10.5$, significantly higher than the usability threshold of 68 (p=$1.24 \times 10^{-5}$), categorizing it as "A" in system usability [7]. In addition, all scores of individual SUS and user experience (UX) questions are significantly better than the neutral score of 3 (p<0.001). The distributions of individual SUS question scores are illustrated in Fig. 4.3. These results indicate positive experience and attitude for various aspects of the proposed system. Specifically, the SUS questionnaire responses highlighted that participants perceived the system as well-integrated (score = $4.6 \pm 0.5$) but expressed lower confidence in task performance (score = $4.0 \pm 0.8$). While they did not find the system complex (score = $1.3 \pm 0.6$), they indicated a preference for technical support (score = $2.3 \pm 1.1$).

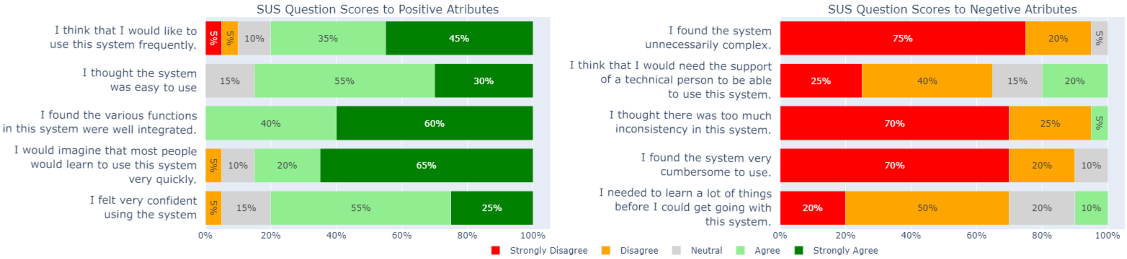


Figure 4.3: Distribution of SUS Question Scores Across Participants.

For the UX questions, all average ratings ranged from 4 to 4.65, with a majority of respondents expressing positive feedback (rating 4 or 5 out of 5) on various aspects. Specifically, 65% found the final visualization informative, 80% found the system ergonomic, 90% felt engaged, and 85% found the hand-grabbing interface and visual feedback for camera placement intuitive. The majority of participants (95%) also found the pivot method for trocar placement intuitive, while 70% found the information panels helpful. The assessments of the individual UX questions are depicted in Fig. 4.4.

In the open-ended questions, 19 out of 20 participants provided positive and negative aspects of the surgical planning system. Most (14/19) found it easy to use and the feedback metrics helpful (7/19). However, two participants noted the semi-transparent materials hindered depth perception,

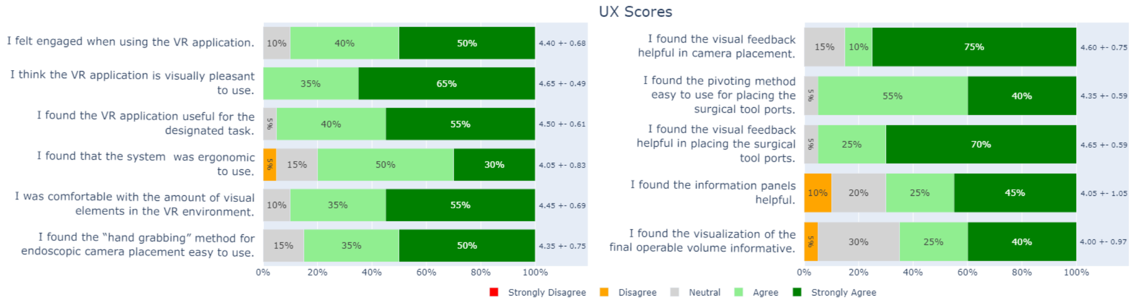


Figure 4.4: Distribution of UX Question Scores across Participants, with mean ± standard deviation displayed beside the respective bar plot.

though visual feedback (White spheres turn into green) helped. Nine participants suggested improvements, e.g., four recommended auditory feedback for guidance and errors, four suggested more guidance for how to optimize surgical planning, such as color-coded manipulation angles on the confirmation panel, and one participant proposed direct 3D model manipulation for height adjustment of the 3D model.

### 4.4.2 Quantitative Evaluation

Trajectory distance, manipulation angle, operable volume, and task completion times were collected from the VR application. In Task 1 (surgical trocar placement), the maximum trajectory distance for both trocars was less than 28 cm, ensuring the surgical target was reachable. The average manipulation angle of 48 degrees was consistent with recommendations from prior research. For Task 2 (camera placement), positioning the camera outside the DOF of other trocars prevented instrument interference and maximized the common area volume, averaging 1 liter of operable volume across participants. We also recorded the number of adjustments and time required for each task during the user study. The summary of these data can be seen in Table 4.1.

The majority of participants (75%) completed both tasks without adjustments. Participants spent an average of $3.70 \pm 1.52$ minutes on planning, with Task 1 taking $1.37 \pm 0.89$ minutes and Task 2 taking $2.33 \pm 1.15$ minutes. Our statistical analysis also revealed significant negative correlations between time spent on the surgical planning and anatomy familiarity (p = 0.041 and correlation = -0.460). This suggests familiarity with the human anatomy can boost performance efficiency.

Table 4.1: Quantitative Evaluation from the User Study

| Task | Metric | Result |
|---|---|---|
| Surgical Trocar placement | Time (Minutes) | 1.37 ± 0.89 |
| | Number of Adjustments | 0.35 ± 0.67 |
| | Manipulation Angle | 48.63 ± 7.39 |
| | Right Hand Trajectory Distance (CM) | 25.13 ± 1.84 |
| | Left Hand Trajectory Distance (CM) | 27.07 ± 0.70 |
| Camera Placement | Time (Minutes) | 2.33 ± 1.15 |
| | Number of Adjustments | 0.35 ± 0.67 |
| | Volume of Common Workable Area (Litres) | 1.01 ± 0.12 |

## 4.5 Discussion

In an earlier version of our system, mirroring standard thoracic surgical procedures, participants were required to position the endoscope camera before placing surgical trocars. However, a pilot study involving four participants revealed the necessity for camera adjustments after trocar placement to mitigate instrument crowding and optimize the shared workspace. Consultation with our expert surgeon led to the decision to reverse the task order in the final system. Although in typical surgical procedures, the camera is placed before surgical trocars to guide following placements, employing semi-transparent materials in our 3D model enables the view of internal anatomies in our system making this sequence unnecessary. By reversing the task order, we eliminated the redundant camera adjustment step and the potential for instrument fighting during camera placement.

In the semi-quantitative evaluation using the SUS questionnaire and customized UX questions showed promising results. While participants generally found the system well-integrated and easy to use, a lack of confidence and a perceived need for technical support emerged. This may be related to the absence of a definitive metric for optimal surgical view and manipulation angles, despite the incorporation of soft metrics to guide trocar placement. The UX questions highlighted a positive user experience overall, with high engagement and perceived usefulness, which are crucial for future clinical adoption. However, information panels and the final operable volume visualization were slightly less well-received than other items in the UX questions. Participants suggested a voice assistant for guidance and error reporting. Notably, those who found the operable volume visualization informative reported lower system complexity and less need for technical support in the SUS questionnaire, resulting in higher overall SUS scores. Regarding the “freehand” camera placement and pivot mechanism, most participants responded favorably and found the visual feedback helpful. Notably, 15% of participants held a neutral view of the freehand camera placement and its feedback, compared to only 5% for surgical trocar placement, suggesting an area for potential improvement. Finally, with a short planning time ($3.70 \pm 1.52$ minutes) with no failed surgical plans, our proposed system offers high efficiency and robustness, required for clinical use.

The current study has several limitations. First, semi-transparent rendering of anatomical structures (e.g., ribs, skin) compromised depth perception. Second, varying difficulty levels for trocar placement based on surgical target location and individual anatomy were not fully explored due to time constraints and the use of one patient model. Third, the limited number of anatomical structures included in the 3D model, due to visualization challenges and computational complexity of segmentation, restricted the development of comprehensive metrics of the proposed system. For example, incorporating the chest wall muscles could help in defining metrics to avoid thick muscles in the chest wall, which can minimize tissue damage and bleeding, while maximizing ease of motion during camera placement. Finally, in our preliminary study, we only recruited non-clinicians for system validation due to the limited accessibility to thoracic surgeons, although the system development greatly benefited from the expertise of our surgical collaborator. Future work will focus on addressing these limitations through alternative visualization techniques, a wider range of patient models, refining the system's metrics and guidelines in collaboration with clinicians, and additional clinical participants in extended system validation upon further refinement.

## 4.6 Conclusion

In this paper, we present the first pre-operative planning VR system designed to optimize trocar placement in thoracic lung surgeries. Our system incorporates an effective pivoting mechanism and a hand-grabbing method, both seamlessly integrated with visual feedback, to help users in the planning process. A comprehensive user study revealed promising results regarding system usability and overall user satisfaction. The insights from the VR system design and assessment can provide important information for similar surgical VR system development, which has a profound potential in clinical practice.

# Chapter 5

# Conclusion and Future Work

## 5.1 Conclusion

This thesis investigated integrating deep learning and a virtual reality environment to improve preoperative planning in minimally invasive thoracic surgery. Recognizing the important role of precise, patient-specific 3D models in surgical planning, the first part of this thesis involved a detailed analysis of various U-shaped deep learning models for segmenting thoracic anatomy from CT images. This analysis evaluated the effects of different network configurations and attention mechanisms, identifying CNN-based architectures as the most effective for creating accurate and efficient segmentation maps. These segmentation maps form the foundation for constructing the patient-specific 3D models essential to the VR-based planning system.

Building on these segmentation capabilities, the second part of this thesis proposed a rule-based VR system for optimizing trocar placement using right upper lobectomy as a case study. This VR environment allowed surgeons to interactively explore and evaluate trocar placement within a patient-specific model, facilitating observation to establish surgical principles and potentially enhancing surgical outcomes. A preliminary user study involving 20 participants confirmed the system's user-friendliness, robustness, and efficiency, highlighting its promise for clinical application.

In summary, this thesis provided valuable insights for selecting optimal deep learning models to generate accurate anatomical models and demonstrated the transformative potential of VR systems in advancing preoperative planning for minimally invasive lung surgery.

## 5.2 Future Work

One potential direction for future work, on the first contribution could involve investigating different skip connection schemes and their impact on model performance. Various skip-connection schemes have been introduced in U-shaped models to enhance accuracy by reducing the semantic gap between encoder and decoder features, enabling more effective feature aggregation. For example, UNet++ [97] used a nested hierarchical skip-connection scheme to refine feature aggregation, while UCTransNet [85] employed transformers between the encoder and decoder to improve feature aggregation. Similarly, BiO-Net [89] incorporates bi-directional skip connections to reuse parameters, further boosting accuracy. Future work could investigate the effects of modifying these architectures by replacing the CNN-based backbones in UNet++ and BiO-Net with Transformer-based or MambaVision-based models and replacing the Transformer in UCTransNet skip-connection with MambaVision. Such experiments could shed light on the effectiveness of skip-connection schemes across different architectures, offering insights that may lead to improved feature aggregation and accuracy in future model development.

Another promising avenue could be optimizing the STUNet architecture, which showed strong performance in our experiments. By incorporating attention mechanisms into its skip connections, STUNet could capture global contextual features. This enhancement would help the model retain important information across layers and improve accuracy.

Finally, future work could explore semi-supervised learning techniques to improve model performance by using unlabeled data during training. Semi-supervised approaches can unlock more capacity within models, improving generalization across datasets. Benchmarking various models in a semi-supervised training pipeline could provide valuable insights into the role of unlabeled data in model accuracy and robustness, helping to refine the training pipeline for optimal performance.

For the VR application described in the second contribution, there are several possibilities for future development. Currently, the patient-specific model is static, and physical tissue response is not simulated. One direction for further work could involve making the patient's 3D model deformable by integrating techniques such as Finite Element Modeling (FEM) to simulate realistic tissue behavior. Another promising direction would be to design optimization algorithms that recommend initial trocar placement on the patient's body. Surgeons could then adjust this placement based on their techniques. A possible approach for this could involve designing a fuzzy logic system based on established principles and guidelines for trocar placement. This fuzzy system could also be extended to evaluate the final score of the trocar placement, providing an additional layer of decision support for surgeons.

Also, the system can include multi-user functionality, allowing the surgical team to practice the procedure together. This feature would be beneficial when the surgeon needs different views during the operation. It would also facilitate the practice of camera placement to achieve those views during simulations. This approach could enhance the assistant's familiarity with the surgical process and potentially reduce fatigue for the surgical team, as well as decrease the overall operation time. Moreover, the patient could observe the surgery and become more familiar with the procedure, which may help reduce their stress leading up to the operation. To provide further information between the 3D model and the preoperative images, the tool's trajectory path can be mapped to the preoperative images and shown within the simulation.

Finally, future work could explore directly integrating user interfaces for deep learning models within the VR system. This would enable the automatic generation of patient-specific 3D models from CT images, which could be loaded directly into the VR environment for further exploration and surgical planning.

Appendix A

# Supplementary Materials for Architecture Analysis and Benchmarking of 3D U-shaped Deep Learning Models for Thoracic Anatomical Segmentation

## A.1 The Distribution of the labels on the Train and Test dataset

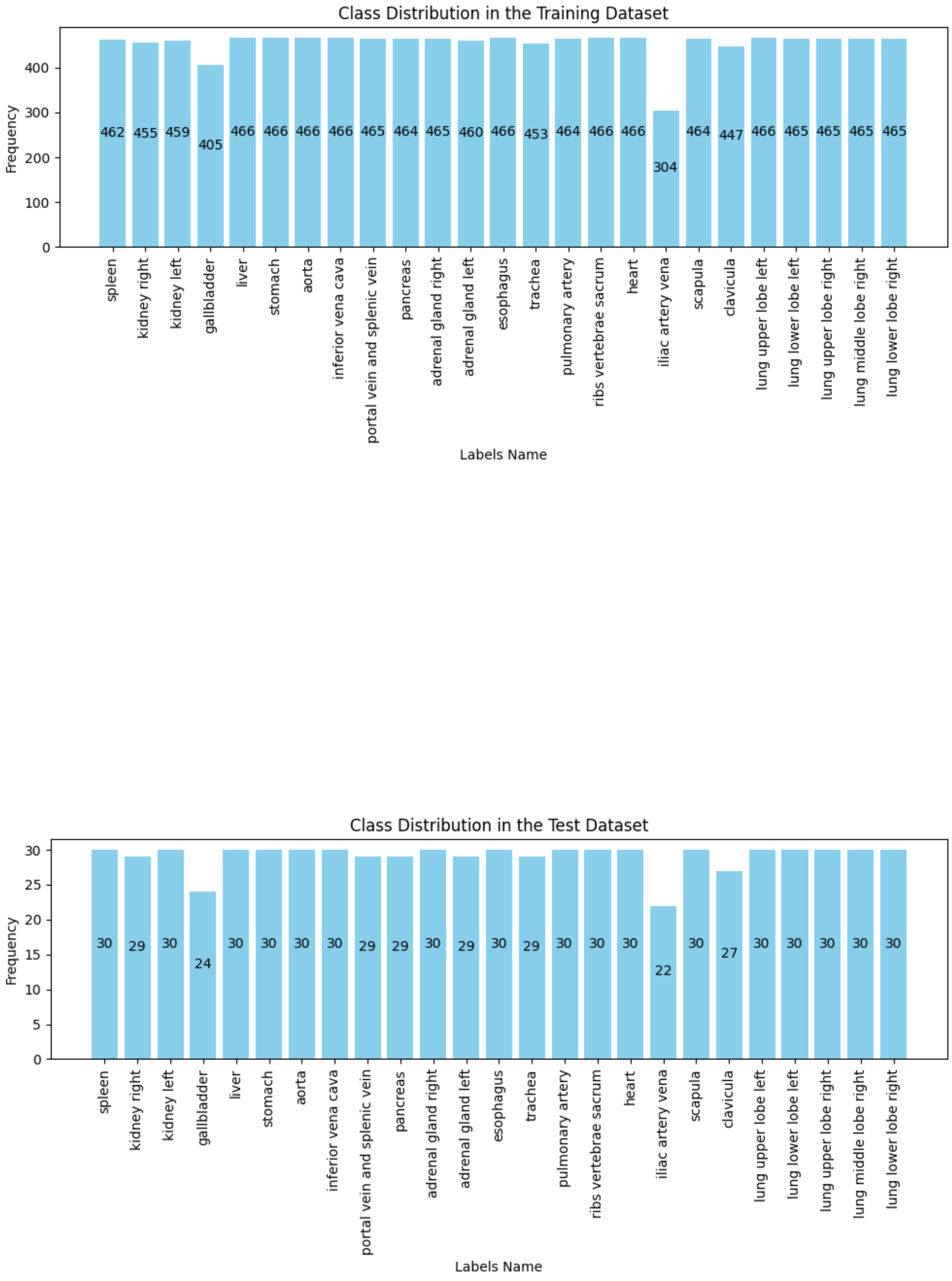


Figure A.1: Frequency distribution for each label class in the training and test datasets.

## A.2 Architectures of the selected U-shaped models

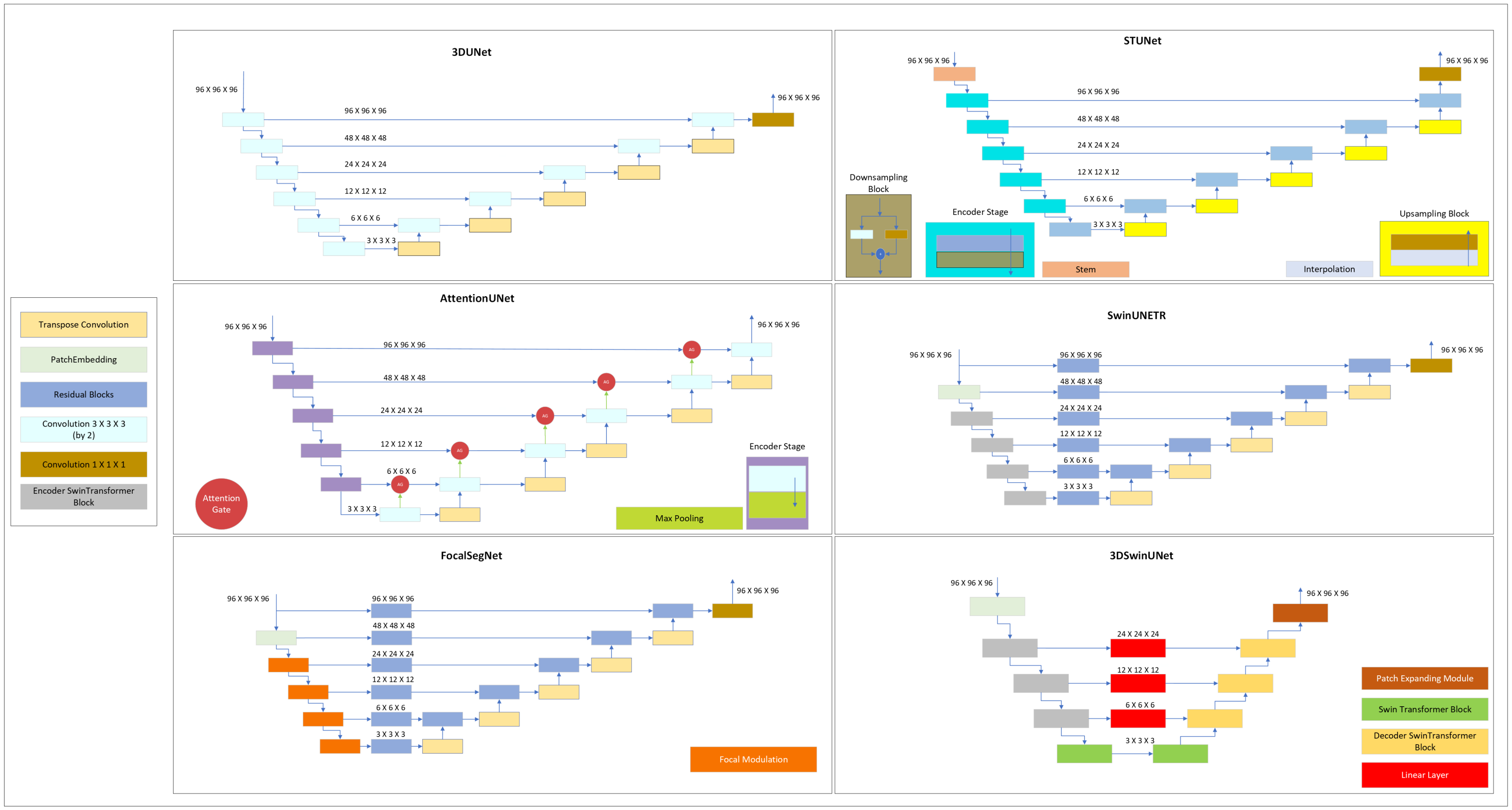


Figure A.2: Model architecture diagrams for the selected U-shaped models.

## A.3 Segmentation results across different models

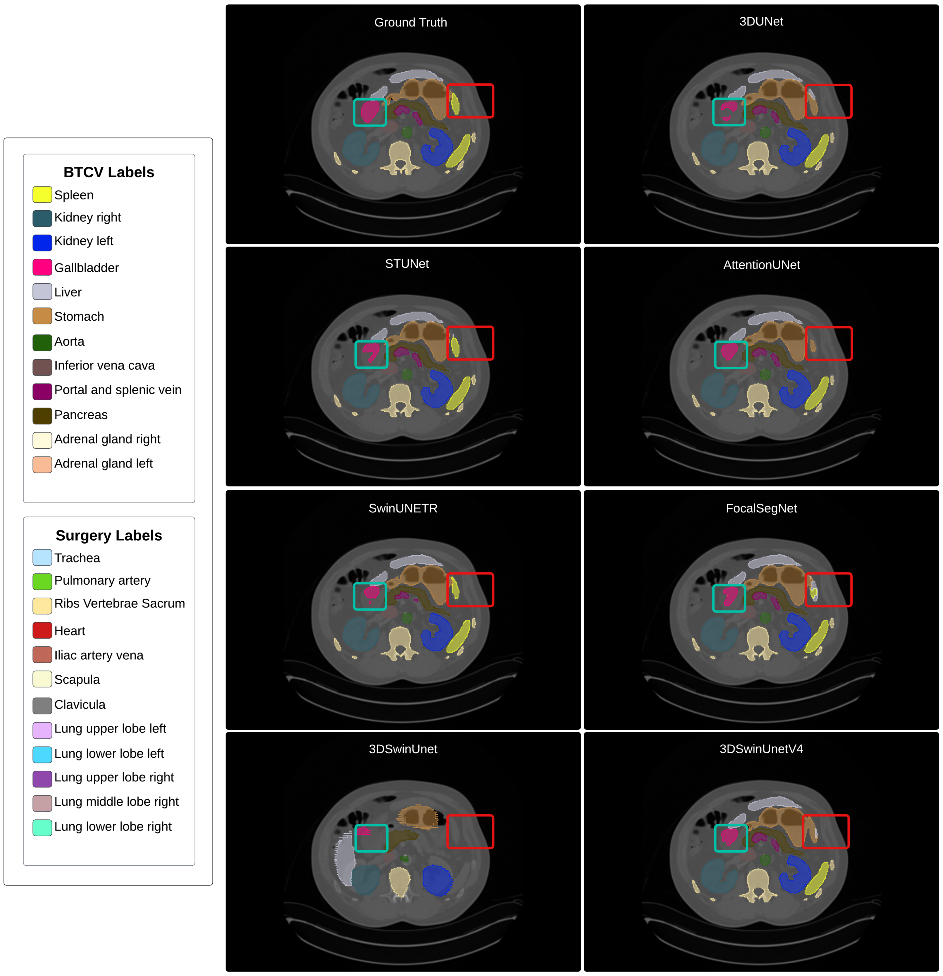


Figure A.3: Visual demonstration of segmentation results with an axial cross-section of a subject's CT scan from different models. The differences across different models are highlighted using the cyan and red boxes.

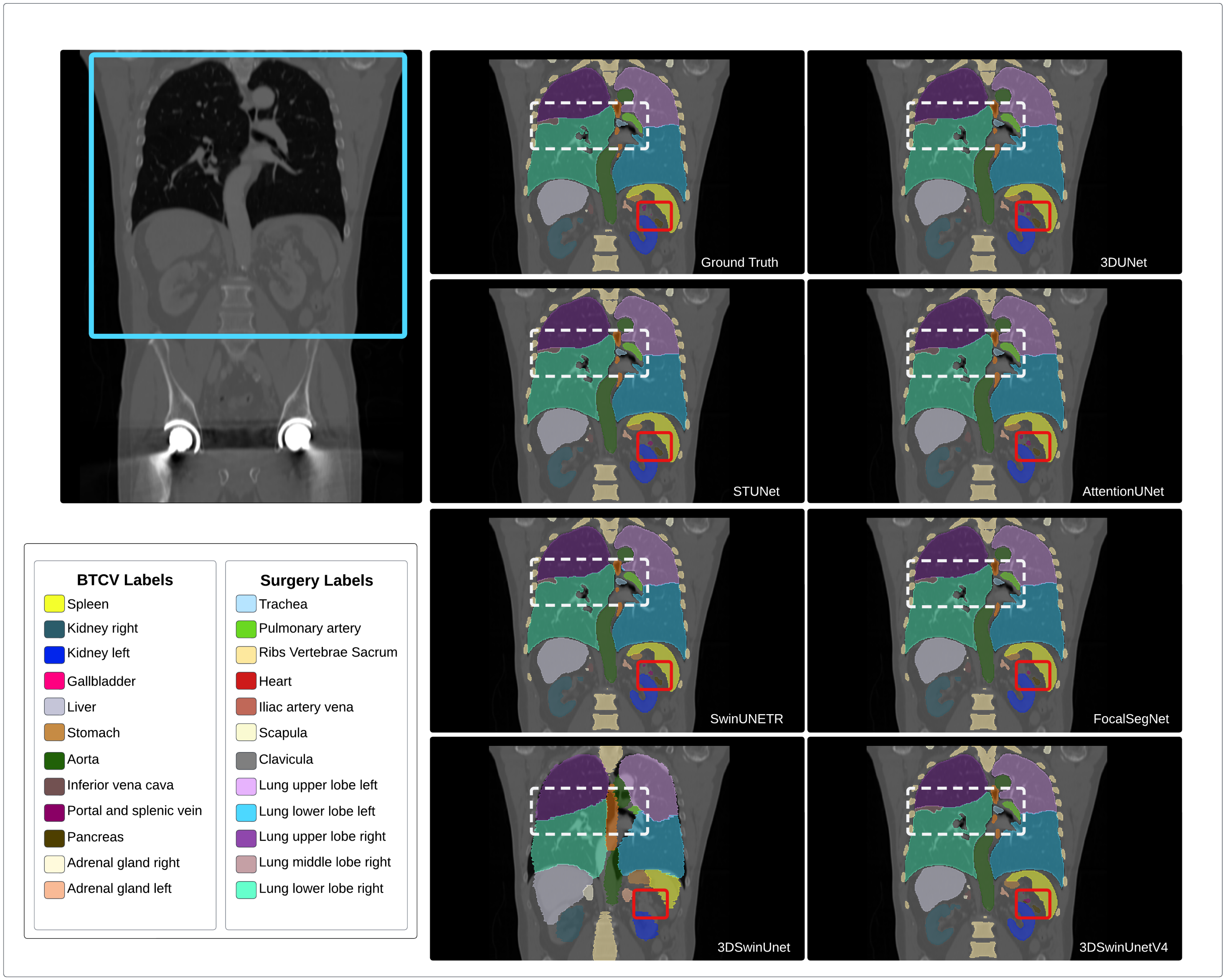


Figure A.4: Visual demonstration of segmentation results with a coronal cross-section of a subject's CT scan from different models. The differences across different models are highlighted using the white and red boxes.

## A.4 Statistical Plots of Segmentation Accuracy

### A.4.1 Boxplots of the Dice Scores among different classes

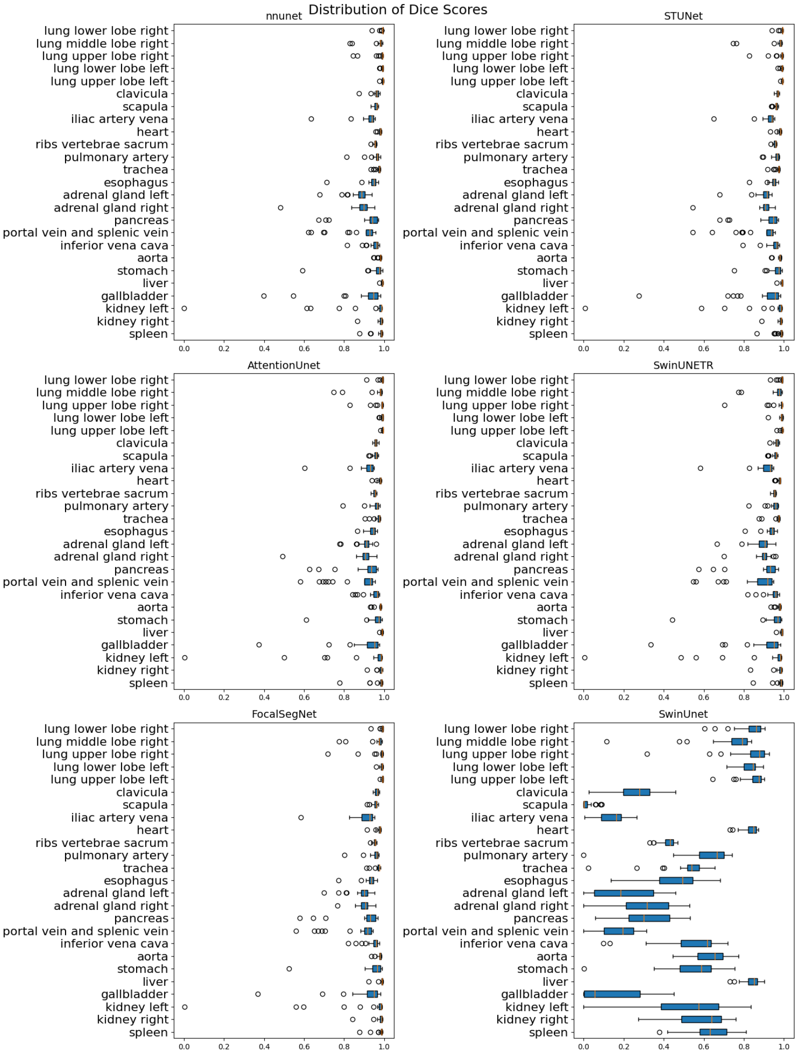


Figure A.5: Boxplots of Dice scores per class across different U-shaped models.

### A.4.2 Distribution of the NSD Score among classes

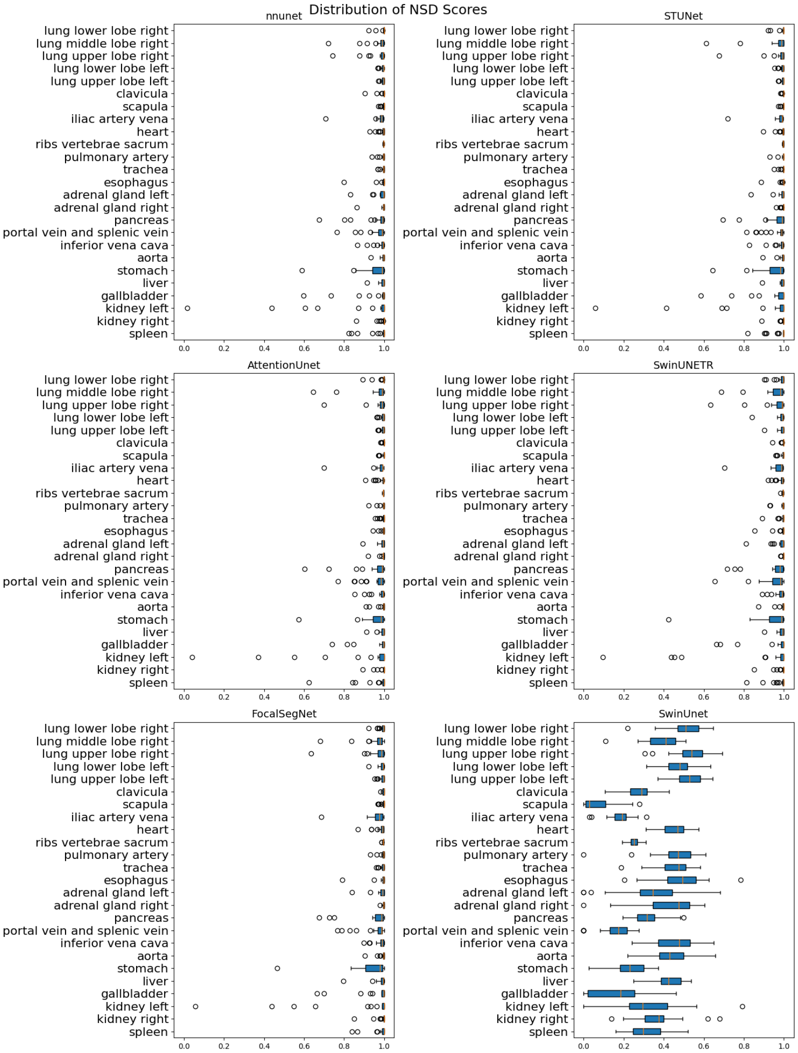
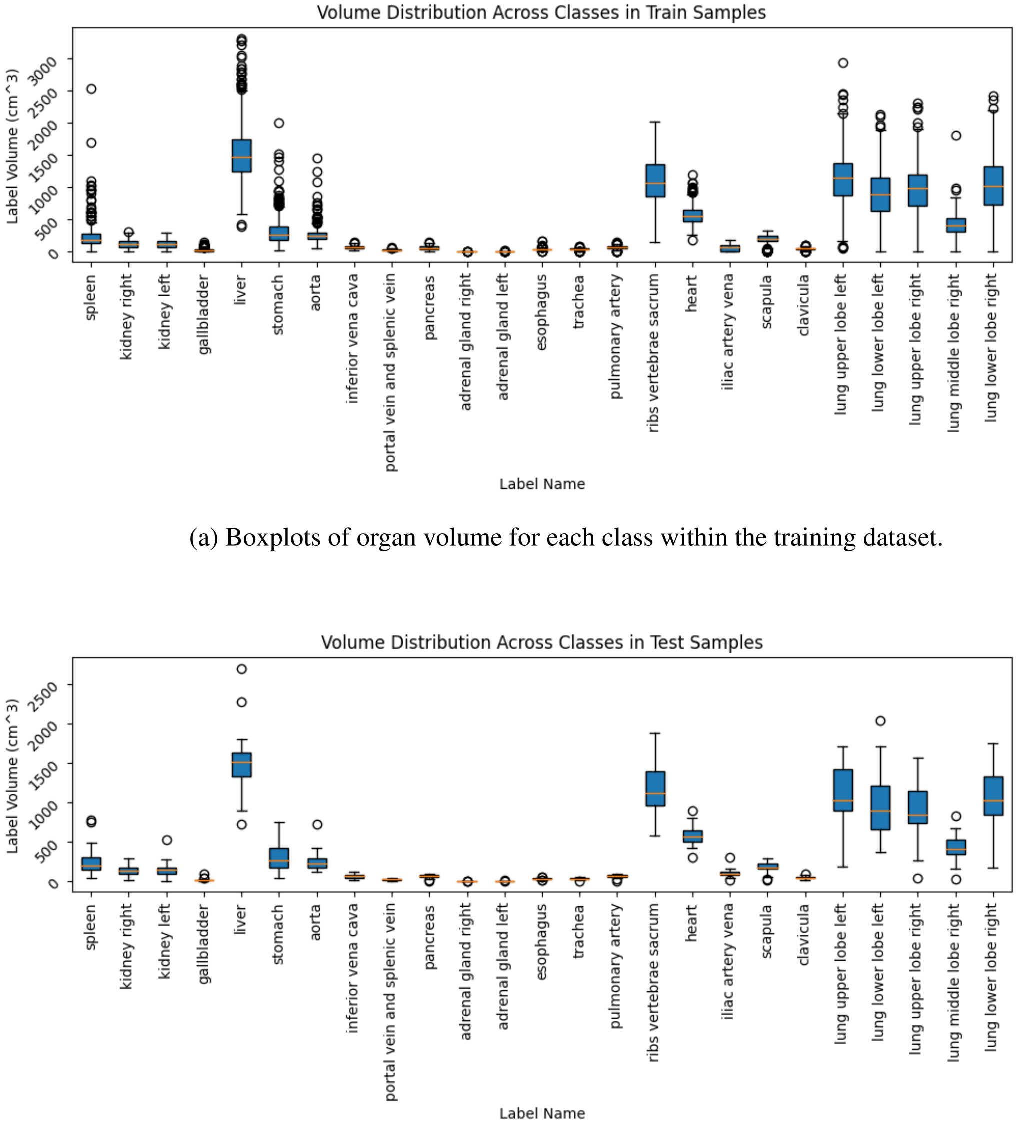


Figure A.6: Boxplots of Normalized Surface Distance (NSD) scores per class across different U-shaped models.

## A.5 The volume distribution of the labels in the Train and Test dataset



(a) Boxplots of organ volume for each class within the training dataset.

(b) Boxplots of organ volume for each class within the test dataset.

Figure A.7: Boxplots of organ volume for each class within the training and test datasets.

## A.6 Jaccard Similarity Metric of the Models

Table A.1: Jaccard Similarity Metric (mean±std) across 3DUNet, STUNet, AttentionUNet, and 3DSwinUnet models with different numbers of resolution stages.

| **Model Name** | **Jaccard** ↑ | | |
|---|---|---|---|
| | **BTCV** | **Surgery** | **Total** |
| $3DUNet$ | $89.10 \pm 4.28$ | $94.28 \pm 1.85$ | $91.48 \pm 2.41$ |
| $STUNet$ | $89.75 \pm 4.04$ | $94.43 \pm 1.98$ | $91.89 \pm 2.31$ |
| $AttentionUNet$ | $89.26 \pm 4.26$ | $93.94 \pm 2.12$ | $91.40 \pm 2.46$ |
| $SwinUNETR$ | $88.59 \pm 4.41$ | $93.54 \pm 2.32$ | $90.86 \pm 2.60$ |
| $FocalSegNet$ | $88.74 \pm 4.10$ | $93.60 \pm 2.23$ | $90.97 \pm 2.38$ |
| $3DSwinUnet$ | $33.19 \pm 5.59$ | $48.06 \pm 4.85$ | $40.05 \pm 4.58$ |
| $3DSwinUNetV4$ | $86.47 \pm 4.72$ | $92.14 \pm 2.47$ | $89.07 \pm 2.74$ |

## Appendix B

# User Study Questionnaire and Ethics Approval of Virtual Reality-Based Preoperative Planning for Optimized Trocar Placement in Thoracic Surgery: A Preliminary Study

## B.1 User Study Questionnaire

### Participant Information

**Study ID:** ______ **Age:** ______

**Date:** ______ **Sex:** ______

### System Usability Scale (SUS)

Please answer the following questions based on the system you just used. Select a value that best describes your experience from 1 to 5 as directed.

(1) I think that I would like to use this system frequently

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(2) I found the system unnecessarily complex

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(3) I thought the system was easy to use

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(4) I think that I would need the support of a technical person to be able to use this system

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(5) I found the various functions in this system were well integrated

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(6) I thought there was too much inconsistency in this system

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(7) I would imagine that most people would learn to use this system very quickly

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(8) I found the system very cumbersome to use

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(9) I felt very confident using the system

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(10) I needed to learn a lot of things before I could get going with this system

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

## User experience questions

(1) I felt engaged when using the VR application.

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(2) I think the VR application is visually pleasant to use.

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(3) I found the VR application useful for the designated task.

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(4) I found that the system was ergonomic to use.

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(5) I was comfortable with the amount of visual elements in the VR environment.

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(6) I found the "hand grabbing" method for endoscopic camera placement easy to use.

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(7) I found the visual feedback (e.g., guidelines, color visual cues, camera port) helpful in camera placement.

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(8) I found the pivoting method easy to use for placing the surgical tool ports.

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(9) I found the visual feedback (e.g., trajectory paths, color visual cues, degree-of-freedom cones) helpful in placing the surgical tool ports.

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(10) I found the information panels (front wall to display instructions and confirmation/warning prompt windows) helpful.

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

(11) I found the visualization of the final operatable volume informative.

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

**Strongly Disagree** **Strongly Agree**

## Additional Information

Please answer the following general questions regarding the system. These questions are a mix of open answer and multiple-choice.

(1) What are the positives and negatives of the system?

__________________________________________________________

(2) Any suggestions on how the system can be improved?

__________________________________________________________

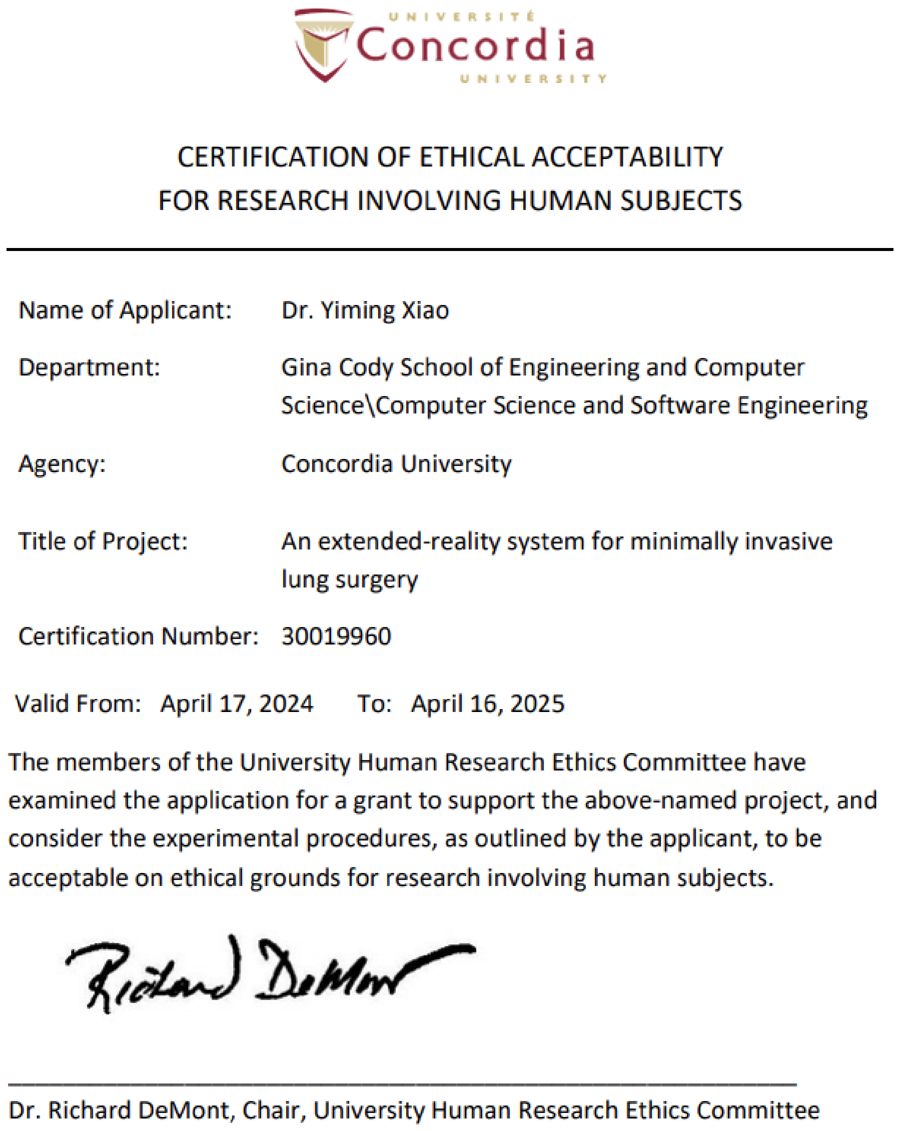
**How familiar were you with Virtual Reality (or other Extended Reality techniques) before performing the study?**

- Familiar
- Somewhat Familiar
- Neutral
- Somewhat Unfamiliar
- Unfamiliar

**How familiar were you with human anatomy before performing the study?**

- Familiar
- Somewhat Familiar
- Neutral
- Somewhat Unfamiliar
- Unfamiliar

## B.2 Ethics Approval Form



UNIVERSITÉ
Concordia
UNIVERSITY

CERTIFICATION OF ETHICAL ACCEPTABILITY
FOR RESEARCH INVOLVING HUMAN SUBJECTS

| | |
|---|---|
| Name of Applicant: | Dr. Yiming Xiao |
| Department: | Gina Cody School of Engineering and Computer Science\Computer Science and Software Engineering |
| Agency: | Concordia University |
| Title of Project: | An extended-reality system for minimally invasive lung surgery |
| Certification Number: | 30019960 |

Valid From: April 17, 2024 To: April 16, 2025

The members of the University Human Research Ethics Committee have examined the application for a grant to support the above-named project, and consider the experimental procedures, as outlined by the applicant, to be acceptable on ethical grounds for research involving human subjects.

Richard DeMont

Dr. Richard DeMont, Chair, University Human Research Ethics Committee

# Bibliography

[1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[2] Wouter Bakhuis, Amir H Sadeghi, Iris Moes, Alexander PWM Maat, Sabrina Siregar, Ad JJC Bogers, and Edris AF Mahtab. Essential surgical plan modifications after virtual reality planning in 50 consecutive segmentectomies. *The Annals of Thoracic Surgery*, 115(5):1247–1255, 2023.

[3] Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Jalen Anthony Benson, Weiruo Zhang, Ann NC Leung, et al. A radiogenomic dataset of non-small cell lung cancer. *Scientific data*, 5(1):1–9, 2018.

[4] R Bauernschmitt, M Feuerstein, J Traub, Eva U Schirmbeck, G Klinker, and R Lange. Optimal port placement and enhanced guidance in robotically assisted cardiac surgery. *Surgical endoscopy*, 21:684–687, 2007.

[5] Morten Bendixen, Ole Dan Jørgensen, Christian Kronborg, Claus Andersen, and Peter Bjørn Licht. Postoperative pain and quality of life after lobectomy via video-assisted thoracoscopic surgery or anterolateral thoracotomy for early stage lung cancer: a randomised controlled trial. *The Lancet Oncology*, 17(6):836–844, 2016.

[6] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester,

et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

[7] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189 (194):4–7, 1996.

[8] Catherine T Byrd, Kiah M Williams, and Leah M Backhus. A brief overview of thoracic surgery in the united states. *Journal of Thoracic Disease*, 14(1):218, 2022.

[9] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[10] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

[11] Jianzheng Cen, Rong Liufu, Shusheng Wen, Hailong Qiu, Xiaobin Liu, Xiaokun Chen, Haiyun Yuan, Meiping Huang, and Jian Zhuang. Three-dimensional printing, virtual reality and mixed reality for pulmonary atresia: early surgical outcomes evaluation. *Heart, Lung and Circulation*, 30(2):296–302, 2021.

[12] Ernest G Chan, James R Landreneau, Matthew J Schuchert, David D Odell, Suicheng Gu, Jiantao Pu, James D Luketich, and Rodney J Landreneau. Preoperative (3-dimensional) computed tomography lung reconstruction before anatomic segmentectomy or lobectomy for stage i non–small cell lung cancer. *The Journal of thoracic and cardiovascular surgery*, 150(3):523–528, 2015.

[13] B Chen, Y Liu, Z Zhang, G Lu, and D Zhang. Transattunet: multi-level attention-guided u-net with transformer for medical image segmentation. arxiv. *arXiv preprint arXiv:2107.05274*, 2021.

[14] Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: a review. *Frontiers in cardiovascular medicine*, 7:25, 2020.

[15] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[16] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016.

[17] CreateXR. The virtuality spectrum: Understanding ar, mr, vr, and xr, 2021. URL `https://creatxr.com/the-virtuality-spectrum-understanding-ar-mr-vr-and-xr/`.

[18] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34:3965–3977, 2021.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[20] Amirreza Fateh, Reza Tahmasbi Birgani, Mansoor Fateh, and Vahid Abolghasemi. Advancing multilingual handwritten numeral recognition with attention-driven transfer learning. *IEEE Access*, 12:41381–41395, 2024.

[21] Marco Feuerstein, Stephen M Wildhirt, Robert Bauernschmitt, and Nassir Navab. Automatic patient registration for port placement in minimally invasixe endoscopic surgery. In *Medical*

*Image Computing and Computer-Assisted Intervention–MICCAI 2005: 8th International Conference, Palm Springs, CA, USA, October 26-29, 2005, Proceedings, Part II 8*, pages 287–294. Springer, 2005.

[22] Marco Feuerstein, Thomas Mussack, Sandro M Heining, and Nassir Navab. Intraoperative laparoscope augmentation for port placement and resection planning in minimally invasive liver resection. *IEEE Transactions on Medical Imaging*, 27(3):355–369, 2008.

[23] Wolfire Games. Triangle mesh voxelization, 2009. URL http://blog.wolfire.com/2009/11/Triangle-mesh-voxelization.

[24] Reena M Ghosh, Matthew A Jolley, Christopher E Mascio, Jonathan M Chen, Stephanie Fuller, Jonathan J Rome, Elizabeth Silvestro, and Kevin K Whitehead. Clinical 3d modeling to guide pediatric cardiothoracic surgery and intervention using 3d printed anatomic models, computer aided design and virtual reality. *3D Printing in Medicine*, 8(1):11, 2022.

[25] Daniel Gut, Zbisław Tabor, Mateusz Szymkowski, Miłosz Rozynek, Iwona Kucybała, and Wadim Wojciechowski. Benchmarking of deep architectures for segmentation of medical images. *IEEE Transactions on Medical Imaging*, 41(11):3231–3241, 2022.

[26] Tamim Ahmad Haidari, Flemming Bjerrum, Henrik Jessen Hansen, Lars Konge, and René Horsleben Petersen. Simulation-based vats resection of the five lung lobes: a technical skills test. *Surgical Endoscopy*, pages 1–9, 2022.

[27] GB Hanna, S Shimi, and A Cuschieri. Optimal port locations for endoscopic intracorporeal knotting. *Surgical Endoscop*, 11:397–401, 1997.

[28] Arash Harirpoush, George Rakovich, Marta Kersten-Oertel, and Yiming Xiao. Virtual reality-based preoperative planning for optimized trocar placement in thoracic surgery: A preliminary study. *arXiv preprint arXiv:2409.04414*, 2024.

[29] Arash Harirpoush, Amirhossein Rasoulian, Marta Kersten-Oertel, and Yiming Xiao. Architecture analysis and benchmarking of 3d u-shaped deep learning models for thoracic anatomical segmentation. *IEEE Access*, 12:127592–127603, 2024. doi: 10.1109/ACCESS.2024.3456674.

[30] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*, pages 272–284. Springer, 2022.

[31] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[32] Andreas Hammer Håversen, Durga Prasad Bavirisetti, Gabriel Hanssen Kiss, and Frank Lindseth. Qt-unet: A self-supervised self-querying all-transformer u-net for 3d segmentation. *IEEE Access*, 2024.

[33] Yufan He, Vishwesh Nath, Dong Yang, Yucheng Tang, Andriy Myronenko, and Daguang Xu. Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 416–426. Springer, 2023.

[34] Lukas Hedegaard. Pytorch-benchmark, 10 2022.

[35] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019.

[36] Samuel Heuts, Peyman Sardari Nia, and Jos G Maessen. Preoperative planning of thoracic surgery with use of three-dimensional reconstruction, rapid prototyping, simulation and virtual navigation. *Journal of Visualized Surgery*, 2, 2016.

[37] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[38] Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, and Yu Qiao. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. *arXiv preprint arXiv:2304.06716*, 2023.

[39] F Isensee, T Wald, C Ulrich, M Baumgartner, S Roy, K Maier-Hein, and PF Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation,(2024). *arXiv preprint arXiv:2404.09556*, 2024.

[40] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[41] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 488–498. Springer, 2024.

[42] Ahmad Jameel Ismail and RK Mishra. Comparing task performance and comfort during non-pulmo nary video-assisted thoracic surgery procedures between the application of the ‘baseball diamond’and the ‘triangle target’principles of port placement in swine models. *World*, 7(2): 60–65, 2014.

[43] Katrine Jensen, Flemming Bjerrum, Henrik Jessen Hansen, René Horsleben Petersen, Jesper Holst Pedersen, and Lars Konge. A new possibility in thoracoscopic virtual reality simulation training: development and testing of a novel virtual reality simulator for video-assisted thoracoscopic surgery lobectomy. *Interactive cardiovascular and thoracic surgery*, 21(4): 420–426, 2015.

[44] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems*, 35:36722–36732, 2022.

[45] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.

[46] Jason Kugelman, Joseph Allman, Scott A Read, Stephen J Vincent, Janelle Tong, Michael Kalloniatis, Fred K Chen, Michael J Collins, and David Alonso-Caneiro. A comparison of deep learning u-net architectures for posterior segment oct retinal layer segmentation. *Scientific reports*, 12(1):14888, 2022.

[47] Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. Segthor: Segmentation of thoracic organs at risk in ct images. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2020.

[48] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12, 2015.

[49] Rodney J Landreneau, Michael J Mack, Stephen R Hazelrigg, Robert D Dowling, Tea E Acuff, Mitchell J Magee, and Peter F Ferson. Video-assisted thoracic surgery: basic technical concepts and intercostal approach strategies. *The Annals of thoracic surgery*, 54(4):800–807, 1992.

[50] James R Lewis. The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction*, 34(7):577–590, 2018.

[51] Liangliang Liu, Jianhong Cheng, Quan Quan, Fang-Xiang Wu, Yu-Ping Wang, and Jianxin Wang. A survey on u-shaped networks in medical image segmentations. *Neurocomputing*, 409:244–258, 2020.

[52] Xiangbin Liu, Liping Song, Shuai Liu, and Yudong Zhang. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13(3):1224, 2021.

[53] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[54] Fernando López-Mir, V Naranjo, JJ Fuertes, M Alcañiz, J Bueno, and E Pareja. Design and validation of an augmented reality system for laparoscopic surgery in a real environment. *BioMed Research International*, 2013(1):758491, 2013.

[55] Shi-ping Luh and Hui-ping Liu. Video-assisted thoracic surgery—the past, present status and the future. *Journal of Zhejiang University Science B*, 7:118–128, 2006.

[56] Mohammad R Maddah, Jean-Marc Classe, Isabelle Jaffre, Keith A Watson, Katherine S Lin, Damien Chablat, Cédric Dumas, and Caroline GL Cao. A decision aid for the port placement problem in robot-assisted hysterectomy. *Laparoscopic, Endoscopic and Robotic Surgery*, 6 (2):43–56, 2023.

[57] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P Bradley, Aaron Carass, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature communications*, 9(1):5217, 2018.

[58] Mattatz. Unity voxel. https://github.com/mattatz/unity-voxel, 2019.

[59] Paul Milgram and Fumio Kishino. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems*, 77(12):1321–1329, 1994.

[60] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[61] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[62] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

[63] Himashi Peiris, Munawar Hayat, Zhaolin Chen, Gary Egan, and Mehrtash Harandi. A robust volumetric transformer for accurate 3d tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 162–172. Springer, 2022.

[64] Alexandra L. Potter, Thrusha Puttaraju, Jon C. Sulit, Jorind Beqari, Camille A. Mathey Andrews, Arvind Kumar, Maya Sharma, Mika Sharma, Phillip J. Spencer, and Chi-Fu Jeffrey Yang. Assessing the number of annual lung cancer resections performed in the united states. *Shanghai Chest*, 7(0), 2023. ISSN 2521-3768. URL https://shc.amegroups.org/article/view/8191.

[65] Silviu Daniel Preda, Cătălin Ciobîrcă, Gabriel Gruionu, Andreea Şoimu Iacob, Konstantinos Sapalidis, Lucian Gheorghe Gruionu, Ştefan Castravete, Ştefan Pătraşcu, and Valeriu Şurlin. Preoperative computer-assisted laparoscopy planning for the minimally invasive surgical repair of hiatal hernia. *Diagnostics*, 10(9):621, 2020.

[66] Qiumei Pu, Zuoxin Xi, Shuai Yin, Zhe Zhao, and Lina Zhao. Advantages of transformer and its application for medical image segmentation: a survey. *BioMedical Engineering OnLine*, 23(1):14, 2024.

[67] Bin Qiu, Ying Ji, Huayu He, Jun Zhao, Qi Xue, and Shugeng Gao. Three-dimensional reconstruction/personalized three-dimensional printed model for thoracoscopic anatomical partial-lobectomy in stage i lung cancer: a retrospective study. *Translational Lung Cancer Research*, 9(4):1235, 2020.

[68] Amirhossein Rasoulian, Soorena Salari, and Yiming Xiao. Weakly supervised segmentation of intracranial aneurysms using a 3d focal modulation unet. *arXiv preprint arXiv:2308.03001*, 2023.

[69] Annika Reinke, Minu D Tizabi, Carole H Sudre, Matthias Eisenmann, Tim Rädsch, Michael Baumgartner, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, et al. Common

limitations of image processing metrics: A picture story. *arXiv preprint arXiv:2104.05642*, 2021.

[70] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[71] Jaiyoung Ryu, William P. Cooney, Linda J. Askew, Kai-Nan An, and Edmund Y.S. Chao. Functional ranges of motion of the wrist joint. *The Journal of Hand Surgery*, 16(3):409–419, 1991. ISSN 0363-5023. doi: https://doi.org/10.1016/0363-5023(91)90006-W. URL https://www.sciencedirect.com/science/article/pii/036350239190006W.

[72] Peyman Sardari Nia, Jules R Olsthoorn, Samuel Heuts, and Jos G Maessen. Interactive 3d reconstruction of pulmonary anatomy for preoperative planning, virtual simulation, and intraoperative guiding in video-assisted thoracoscopic lung surgery. *Innovations*, 14(1):17–26, 2019.

[73] Masato Sasaki, Seiya Hirai, Masakazu Kawabe, Takahiko Uesaka, Kouichi Morioka, Akio Ihaya, and Kuniyoshi Tanaka. Triangle target principle for the placement of trocars during video-assisted thoracic surgery. *European journal of cardio-thoracic surgery*, 27(2):307–312, 2005.

[74] Lovis Schwenderling, Florian Heinrich, and Christian Hansen. Augmented reality visualization of automated path planning for percutaneous interventions: a phantom study. *International Journal of Computer Assisted Radiology and Surgery*, 17(11):2071–2079, 2022.

[75] Manuel Simoes and Caroline GL Cao. Leonardo: A first step towards an interactive decision aid for port-placement in robotic surgery. In *2013 IEEE international conference on systems, man, and cybernetics*, pages 491–496. IEEE, 2013.

[76] Richard Skarbez, Missie Smith, and Mary C Whitton. Revisiting milgram and kishino's reality-virtuality continuum. *Frontiers in Virtual Reality*, 2:647997, 2021.

[77] Brian Solomon, Costas Bizekis, Sophia L Dellis, Jessica S Donington, Aaron Oliker, Leora B Balsam, Michael Zervos, Aubrey C Galloway, Harvey Pass, and Eugene A Grossi. Simulating video-assisted thoracoscopic lobectomy: a virtual reality cognitive task simulation. *The Journal of thoracic and cardiovascular surgery*, 141(1):249–255, 2011.

[78] Igor Stenin, Stefan Hansen, Meike Becker, Georgios Sakas, Dieter Fellner, Thomas Klenzner, and Jörg Schipper. Minimally invasive multiport surgery of the lateral skull base. *BioMed research international*, 2014(1):379295, 2014.

[79] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[80] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[81] Jie Tian, Kaijie Wu, Kai Ma, Hao Cheng, and Chaocheng Gu. Exploration of different attention mechanisms on medical image segmentation. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part IV 26*, pages 598–606. Springer, 2019.

[82] Hideki Ujiie, Aogu Yamaguchi, Alexander Gregor, Harley Chan, Tatsuya Kato, Yasuhiro Hida, Kichizo Kaga, Satoru Wakasa, Chad Eitel, Tod R Clapp, et al. Developing a virtual reality simulation system for preoperative planning of thoracoscopic thoracic surgery. *Journal of Thoracic Disease*, 13(2):778, 2021.

[83] Hideki Ujiie, Ryohei Chiba, Aogu Yamaguchi, Shunsuke Nomura, Haruhiko Shiiya, Aki Fujiwara-Kuroda, Kichizo Kaga, Chad Eitel, Tod R Clapp, and Tatsuya Kato. Developing a virtual reality simulation system for preoperative planning of robotic-assisted thoracic surgery. *Journal of Clinical Medicine*, 13(2):611, 2024.

[84] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[85] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2441–2449, 2022.

[86] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 109–119. Springer, 2021.

[87] Jakob Wasserthal, Manfred Meyer, Hanns-Christian Breit, Joshy Cyriac, Shan Yang, and Martin Segeroth. Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. *arXiv preprint arXiv:2208.05868*, 2022.

[88] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023.

[89] Tiange Xiang, Chaoyi Zhang, Xinyi Wang, Yang Song, Dongnan Liu, Heng Huang, and Weidong Cai. Towards bi-directional skip connections in encoder-decoder architectures and beyond. *Medical Image Analysis*, 78:102420, 2022.

[90] Hanguang Xiao, Li Li, Qiyuan Liu, Xiuhong Zhu, and Qihang Zhang. Transformers in medical image segmentation: A review. *Biomedical Signal Processing and Control*, 84:104791, 2023.

[91] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 171–180. Springer, 2021.

[92] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022.

[93] Wenjian Yao, Jiajun Bai, Wei Liao, Yuheng Chen, Mengjuan Liu, and Yao Xie. From cnn to transformer: A review of medical image segmentation models. *arXiv preprint arXiv:2308.05305*, 2023.

[94] Rammah Yousef, Shakir Khan, Gaurav Gupta, Tamanna Siddiqui, Bader M Albahlal, Saad Abdullah Alajlan, and Mohd Anul Haq. U-net-based models towards optimal mr brain image segmentation. *Diagnostics*, 13(9):1624, 2023.

[95] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021.

[96] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 2023.

[97] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.