Continual Learning in Interactive Medical Image Segmentation

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science Engineering

by

Kushal Borkar
2020701033
kushal.borkar@research.iiit.ac.in

Advisor: Prof. (Dr.) C V Jawahar & Prof. (Dr.) Chetan Arora

International Institute of Information Technology Hyderabad
500 032, India

Jan 2025

International Institute of Information Technology Hyderabad
Hyderabad, India

CERTIFICATE

This is to certify that work presented in this thesis proposal titled ***Continual Learning in Interactive Medical Image Segmentation*** by *Kushal Borkar* has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Prof. C. V. Jawahar, Prof. Chetan Arora

Acknowledgment

First and foremost, I extend my deepest gratitude to my advisors, Prof. Chetan Arora and Prof. C.V. Jawahar, for their unwavering support, insightful guidance, and constant encouragement throughout this intellectually rigorous yet immensely rewarding journey. Their mentorship has been instrumental in shaping my research trajectory, providing me with both the theoretical foundation and the practical direction necessary to navigate the complexities of interactive segmentation. They introduced me to this fascinating field, guiding me through its core principles while helping me identify and refine critical challenges that needed addressing. Their thought-provoking discussions, keen insights, and meticulous feedback were pivotal in shaping my problem statement, ensuring that my work maintained theoretical depth and practical relevance. Their rigorous yet supportive approach has not only elevated the quality of my research but also deeply influenced my broader perspective on scientific inquiry, critical thinking, and innovation. I am profoundly grateful for their patience, wisdom, and the invaluable lessons they have imparted—lessons that extend far beyond the scope of this thesis.

My dedicated project partner, Abhilaksh Singh Reen, deserves special recognition for his invaluable contributions, particularly in developing the software and the refinement of the user interface. His relentless effort in integrating open-source tools like 3D Slicer, and OHF Viewer significantly enhanced the usability and accessibility of our system. His keen eye for detail and deep understanding of user-centric design ensured that the interface was efficient. Our countless brainstorming sessions on improving the tool were not just about functionality—they also sparked deeper discussions about bridging the gap between technical research and practical usability. His persistence, technical expertise, and commitment to excellence were instrumental in shaping the interactive experience of our system, and I truly appreciate his contributions.

My friends, Yash, Piyush, and Rupak, have been more than just research companions—they have been my unwavering pillars of support throughout this journey. From moments of sheer frustration when I grappled with seemingly insurmountable challenges to the breakthroughs when my approach finally started working, they stood by me with unwavering encouragement. Our shared passion for research led to countless late-night brainstorming sessions, often accompanied by spontaneous pizza runs or meals at DLF Street, where our conversations seamlessly transitioned from debugging code and debating research ideas to discussing life's uncertainties and aspirations. Their unique perspectives continuously challenged me to revisit the fundamentals, refine my approach, and push beyond my intellectual limits. Beyond just academic discussions, their ability to provide clarity in moments of doubt, their patience in helping me recognize and rectify mistakes, and their persistent belief in my work helped me navigate

setbacks with resilience. Their friendship transformed what could have been a solitary, arduous research journey into an enriching, collaborative experience that I will forever cherish.

I express my heartfelt gratitude to my family, especially my parents, for their unwavering support, patience, and encouragement throughout this journey. Their belief in me has been a constant source of strength, providing me with the resilience to navigate the challenges of research. Their unconditional love and sacrifices have played an indispensable role in enabling me to pursue my academic aspirations. This work would not have been possible without their steadfast presence and unwavering faith in my abilities.

Abstract

Automated segmentation of medical image volumes holds immense potential to significantly reduce the time and effort required from medical experts for annotation. However, leveraging machine learning for this task remains a formidable challenge due to variations in imaging modalities, the inherent complexity of medical images, and the limited availability of labeled patient data. While existing interactive segmentation methods and foundational models incorporate user-provided prompts to iteratively refine segmentation masks, they often fail to learn the continuity and inter-related information between consecutive slices in a 3D medical image volume. This limitation leads to inconsistencies, spatial discontinuities, and loss of anatomical coherence, ultimately affecting the reliability of segmentation results in clinical applications.

The work proposes a novel interactive segmentation framework that dynamically updates model parameters during inference using a test-time training paradigm guided by user-provided scribbles. Unlike traditional approaches, our method preserves crucial spatial and contextual information from both previously processed slices within the same medical volume and the training dataset through a student-teacher learning mechanism. By leveraging sequential dependencies across slices, our approach ensures smoother and more anatomically consistent segmentation masks while integrating prior knowledge from the training distribution.

We extensively evaluated our framework across diverse datasets, encompassing CT, MRI, and microscopic cell images, demonstrating its superior performance in both efficiency and accuracy. Our method significantly reduces user annotation time by a factor of 6.72× compared to manual annotation workflows and factor of 1.93× compared to state-of-the-art interactive segmentation methods. Furthermore, when benchmarked against foundational segmentation models, our framework achieves a Dice score of 0.9 within just 3–4 user interactions—substantially improving upon the 5–8 interactions required by existing models. This reduction in required interactions translates to a more streamlined and intuitive annotation process for volumetric CT and MRI scans.

Additionally, our framework exhibits strong generalization capabilities, effectively segmenting unseen objects with minimal user guidance while maintaining spatial continuity across slices. Its ability to dynamically adapt during inference, integrate sequential information, and leverage interactive feedback makes it a highly effective tool for medical image analysis.

To facilitate adoption and further research, we will publicly release the full source code, pre-trained models, and the developed annotation tool upon publication. This work underscores the transformative

potential of our framework in enhancing the efficiency, accuracy, and consistency of medical image segmentation, addressing a critical need in the medical imaging field.

Contents

List of Figures

List of Tables

List of Related Publications

[P1] Kushal Borkar, Abhilaksh Singh Reen, C.V. Jawahar, Chetan Arora, **"No Prompting Frozen Foundation Models: Interactive Medical Volume Segmentation using Continual Test Time Adaptation of Compact Models"**, in proceedings of *Proceedings of the Fifteenth Indian Conference on Computer Vision Graphics and Image Processing, ICVGIP*, 2024.

Chapter 1

Introduction

1.1 Motivation

In the field of medicine, CT scans and MRI are two indispensable imaging modalities that offer profound insights into the human body. These advanced tools have become fundamental for healthcare professionals in diagnosing, planning treatments, and monitoring the progression of various medical conditions with remarkable precision. CT scans utilize a series of X-rays to create detailed cross-sectional images of internal structures, making them highly effective for identifying bone fractures, detecting internal bleeding, and assessing the location, size, and condition of tumors. Beyond diagnosis, CT imaging is instrumental in guiding complex surgical procedures and evaluating the efficacy of ongoing treatments [1]. In contrast, MRI employs powerful magnets and radio waves to produce high-resolution images of soft tissues, such as the brain, spinal cord, heart, and liver, offering unparalleled clarity of internal structures. This non-invasive modality has proven invaluable in cardiology, particularly for analyzing the structure and function of the heart and blood vessels, aiding in managing cardiovascular diseases [2].

Clinicians extensively use both imaging techniques to provide a more comprehensive understanding of patient conditions. One of the primary benefits of 3D imaging, especially in cardiology, is its ability to precisely map the heart's spatial orientation, structure, size, function, and blood flow dynamics. These capabilities significantly enhance diagnosing and managing congenital heart defects, valvular disorders, and other cardiovascular abnormalities [3, 4]. Similarly, in orthopedic surgery, 3D imaging facilitates precise visualization of bones, joints, and surrounding soft tissues, enabling surgeons to plan and execute procedures more accurately, thereby minimizing risks and improving patient outcomes [5, 6].

Despite the immense utility of these imaging modalities, the valuable information embedded within medical images often remains hidden at the pixel or voxel level, requiring sophisticated image analysis techniques to extract meaningful insights [7, 8]. One of the most critical and challenging tasks in this domain is image segmentation, which involves delineating specific anatomical structures or regions of interest (ROI) within medical images. This process underpins many diagnostic applications, including shape analysis, volume measurement, and longitudinal tracking of disease progression [9, 10]. Medical Image Segmentation (MIS) is a cornerstone of medical image processing, as it enables the automated ex-

traction of clinically relevant features, such as tumors or organ boundaries, at a granular level. However, this task is fraught with unique challenges, including the variability in anatomical structures, differences in patient anatomy, limited availability of annotated data, and inconsistent imaging responses to contrast agents. Accurate segmentation is crucial for applications like tumor tracking or measuring brain tissue changes over time. Still manual segmentation by experts is labor-intensive, time-consuming, and prone to variability [11, 12]. Moreover, the complexity of medical image segmentation surpasses that of natural image segmentation due to the presence of small, subtle features, such as lesions, that are critical for accurate diagnosis.

Deep learning has revolutionized medical image segmentation, offering advanced solutions that significantly enhance the speed, accuracy, and scalability of image analysis in clinical settings. These models automate the segmentation process, alleviating the workload of radiologists and enabling the rapid analysis of large and complex datasets, which would otherwise be time-consuming and labor-intensive [13, 14]. By leveraging deep learning, medical professionals can reconstruct 3D structures from 2D slices or projections obtained from CT and MRI scans. This allows for precise localization, characterization, and quantification of pathologies such as tumors. This capability improves diagnostic accuracy and facilitates more effective treatment planning. For example, these models have demonstrated remarkable promise in the early detection of diseases like breast cancer by identifying malignant tumors that might otherwise go unnoticed, even by experienced radiologists [15]. Additionally, well-trained models seamlessly integrate 2D and 3D processing workflows, reducing the risk of human error and enhancing clinical precision [13].

Convolutional Neural Networks (CNNs) have long served as a widely adopted approach for medical image segmentation, largely due to their ability to hierarchically extract features through successive linear and nonlinear transformations. Traditional CNNs typically rely on two-dimensional kernels to process image data, but more recent developments have incorporated three-dimensional kernels to better exploit the volumetric nature of medical imaging data. This advancement allows for a more comprehensive analysis by capturing spatial context across multiple dimensions, thereby enhancing analytical capabilities [16].

Despite these improvements, the overall significance of CNNs in the domain of medical image segmentation is increasingly being reconsidered. Efforts to fine-tune Fully Convolutional Networks (FCNs) have revealed intrinsic limitations, particularly in the utilization of intermediate features. The symmetric encoder-decoder architecture characteristic of many FCNs struggles to effectively capture and merge multi-scale information. This results in suboptimal information flow and ultimately, less precise segmentation outcomes. Consequently, while CNNs have played a pivotal role in advancing medical image segmentation, their limitations have spurred the exploration of alternative architectures and hybrid approaches that aim to overcome these challenges [17].

U-Net stands out for its simple yet effective architecture among the most widely used medical image segmentation models. It consists of two main components: an encoder and a decoder, connected by skip connections that enable the model to recover spatial information lost during down-sampling [18].

The original U-Net was designed to process 2D images and generate segmentation maps with the required number of classes. Later, the 3D U-Net [19] was introduced, extending the model to handle volumetric data for applications in 3D medical imaging. Another notable model, V-Net, was specifically designed for volumetric medical image segmentation. It employs a larger kernel size of 5x5x5 and replaces traditional convolutional layers with residual blocks, improving feature extraction and model performance [20].

The development and optimization of segmentation models require extensive expertise and experience. Addressing this challenge, nnU-Net provides a self-configuring network architecture capable of automating the segmentation pipeline across diverse datasets and tasks [21]. This automation facilitates broader application and adaptability, making nnU-Net a valuable advancement in medical image segmentation.

Despite their transformative potential, these models are not without challenges. They can occasionally produce false negatives, where critical conditions are overlooked, or false positives, which may lead to unnecessary medical interventions and anxiety for patients. To address these risks, medical AI systems must adhere to stringent accuracy standards and undergo rigorous regulatory scrutiny to ensure reliability and safety in real-world clinical applications [22]. Integrating deep learning into medical image segmentation has not only unlocked new opportunities for early disease detection and personalized treatment but has also made it a vital area of ongoing research and development. A significant challenge in medical image segmentation is the limited availability of large-scale annotated datasets. Publicly available datasets are often small, while more extensive datasets are typically held by hospitals and private institutions, usually inaccessible due to privacy concerns and the sensitive nature of medical data [15, 23]. Medical imaging is inherently expensive, requiring specialized equipment and trained personnel. Additionally, annotating medical datasets with ground truth labels, such as delineating tumor boundaries or segmenting organs, is labor-intensive. Medical experts typically perform this task, making it both time-consuming and costly [23]. These limitations have driven the demand for more generalizable semi-supervised or unsupervised deep learning models that can achieve accurate results with minimal reliance on annotated data, thereby addressing the bottlenecks associated with data acquisition and annotation.

Deep interactive segmentation offers a promising solution to address the trade-off between achieving high-quality segmentation and mitigating the labor-intensive nature of manual annotation. This approach enhances annotation efficiency and accuracy by incorporating human feedback into the training or application process of segmentation models. Users can refine or correct the model output using intuitive inputs such as clicks, scribbles, or fine-grained voxel masks. This interactive process effectively guides the model toward the desired segmentation output, significantly reducing the effort and time required for manual annotations while maintaining high precision. The integration of this approach is particularly valuable in medical image segmentation, where it combines the strengths of automated processes with human expertise, thus improving the overall workflow and enabling more accurate and reliable results.

Building upon this concept, recent deep-learning based techniques [24–28] have increasingly utilized human interaction to enhance segmentation accuracy, particularly in cases where labeled data exists, but fully automatic methods fail to provide precise segmentations. ConvNet-based methods have explored various interaction mechanisms, including bounding boxes [29, 30], polygons [31–33], clicks [30, 34–41], and scribbles [28, 42, 43], to refine segmentation outputs. For example, DEXTR [44] isolates target objects by specifying their four extreme points (left-most, right-most, top-most, and bottom-most pixels), while FCA-Net [34] underscores the critical role of the initial click in improving segmentation performance. More recently, Vision Transformers (ViTs) have emerged as a powerful tool in interactive segmentation. FocalClick [36], leveraging SegFormer [45] as its backbone, achieves state-of-the-art segmentation results with high computational efficiency. Similarly, iSegFormer [37] utilizes the Swin Transformer [46] as a backbone specifically for interactive segmentation in medical imaging. Beyond backbone architectures, other works focus on refinement modules that enhance segmentation quality. For instance, FocusCut [38] and FocalClick [36] introduce local refinement modules designed to improve segmentation precision. However, a notable limitation of these methods is their reliance on domain-specific solutions, as they assume that the model will consistently perform the same segmentation tasks for which it was initially trained, limiting their generalizability to broader applications.

Despite advancements in interactive segmentation techniques, many methods still struggle to capture fine-grained details in user-annotated areas, which is particularly critical in medical image segmentation. A major limitation of current approaches lies in their dependence on static datasets, where all training data is pre-defined, fully available at the outset, and remains unchanged over time. Additionally, these models are typically designed to address a fixed set of tasks or segmentation classes, making them ill-suited for the dynamic and complex nature of real-world medical imaging scenarios. Such environments are often characterized by shifting data distributions, evolving clinical requirements, and diverse imaging modalities.

Furthermore, the performance of many segmentation methods declines when applied to datasets from unseen domains. Addressing this issue, approaches such as IA+SA [47] and RAIS [48] have explored test-time adaptation, where models are updated based on user annotations that provide strong cues about the ground truth. These methods have demonstrated promising results for adapting to single unseen datasets. However, the target domain is rarely stationary in practical large-scale medical image annotation scenarios. Medical datasets often span diverse imaging equipment, modalities, and patient populations, requiring models to adapt to a wide variety of domain-specific characteristics. In such settings, catastrophic forgetting becomes a significant challenge, as adaptation methods risk overwriting previously learned knowledge when exposed to new datasets [47, 48]. To address this, it is imperative to develop strategies that enable models to incrementally learn and accumulate knowledge from new datasets without compromising their performance on previously annotated domains. Such advancements are essential for ensuring robust and generalizable medical image segmentation solutions in real-world clinical applications.

1.2 Context of current research with previous work

This work is a continuation of my previous research on interactive medical image segmentation, building upon the foundation established in *Efficient and Generic Interactive Segmentation Framework to Correct Mispredictions during Clinical Evaluation of Medical Images*. While the prior work focused on correcting segmentation errors, this study advances the field by introducing a novel continual adaptation strategy at test time, enabling the model to refine its predictions dynamically based on user interactions.

A key contribution of this work is the development of a test-time training framework that leverages user-provided scribbles to adapt the model in real-time. Unlike conventional interactive segmentation methods that rely solely on static pre-trained models, our approach continuously learns from each user interaction, improving segmentation accuracy for the current volume while maintaining anatomical consistency across slices. This continual adaptation mechanism integrates a student-teacher learning paradigm, preserving essential prior knowledge while incorporating new information from the ongoing segmentation task.

Through extensive experimentation on diverse medical imaging datasets, including CT, MRI, and microscopic cell images, we demonstrate that our method significantly enhances segmentation accuracy while reducing the number of user interactions required. By leveraging inter-slice continuity and adapting to new data at inference time, our framework achieves superior performance over existing state-of-the-art approaches, making it a more efficient and reliable tool for clinical applications.

1.3 Contributions

This thesis introduces a novel interactive segmentation framework that integrates continual test-time adaptation (TTA) with a teacher-student learning paradigm, allowing dynamic refinement of segmentation models based on user interactions. The primary contributions of this work are as follows:

- **Continual Test-Time Adaptation for Interactive Segmentation:** We propose a novel framework that enables pre-trained segmentation models to adapt dynamically at inference time without requiring any architectural modifications. Our approach refines segmentation predictions in real-time while preserving knowledge from previously segmented slices by leveraging continual test-time training guided by user-provided scribbles.

- **Teacher-Student Learning for Knowledge Preservation:** The proposed method incorporates a teacher-student paradigm, where the student model undergoes test-time adaptation based on user feedback and pseudo-segmentation masks, while the teacher model is updated through an exponential moving average (EMA) of the student's parameters. This design effectively balances adaptation to the current slice while preserving essential knowledge from previous slices and mitigating overfitting issues commonly encountered in naive adaptation strategies.

- **Integration of Sequential Information for Volume Consistency:** Unlike traditional interactive segmentation models that treat each slice independently, our framework explicitly models inter-slice dependencies in volumetric medical images. By incorporating knowledge from both past interactions and adjacent slices, our method ensures anatomical consistency across the segmented volume and addresses a critical limitation of existing segmentation techniques.

- **Flexible and Extensible Architecture:** The proposed framework is designed to be model-agnostic, allowing seamless integration with various segmentation backbones, including transformer-based and convolutional models. This adaptability ensures that our approach remains compatible with future advancements in deep learning-based segmentation models.

- **State-of-the-Art Performance in Interactive Medical Image Segmentation:** Through extensive experimentation on multiple publicly available medical imaging datasets, including CT, MRI, and microscopic cell images, we demonstrate that our method surpasses existing interactive image segmentation techniques. Our framework consistently achieves higher segmentation accuracy while significantly reducing user annotation effort compared to conventional interactive segmentation methods and prompting-based foundational models.

- **Efficient and User-Friendly Interactive Annotation Workflow:** By minimizing the number of user interactions required for high-quality segmentation, our framework substantially reduces annotation time and effort, making it highly suitable for clinical applications. The ability to refine segmentations dynamically with minimal user input enhances its practicality for real-world medical image analysis tasks.

In summary, this thesis advances the field of interactive medical image segmentation by introducing a continual test-time adaptation strategy that enables real-time refinement of segmentation models while maintaining inter-slice consistency. The proposed framework represents a significant step toward improving the efficiency, accuracy, and usability of interactive segmentation techniques in clinical and research settings.

1.4 Structure of the proposal

This thesis is structured into six chapters, each addressing a key aspect of the research on continual test-time adaptation for interactive medical image segmentation. Below is an overview of each chapter:

Chapter 1: Introduction This chapter provides a comprehensive introduction to the research problem, outlining the motivation behind the study and its significance in the field of medical image segmentation. It begins by discussing the challenges in interactive medical image segmentation and the need for continual adaptation at test time. The chapter then establishes the context of the current research in relation to previous work, highlighting the advancements made in this study. It also presents the key contributions of this research and concludes with an overview of the thesis structure.

Chapter 2: Background This chapter provides the necessary background knowledge required to understand the research problem. It starts with an overview of medical image segmentation and its importance in clinical applications. Different medical imaging modalities, such as Radiography (X-ray), Magnetic Resonance Imaging (MRI), and Computed Tomography (CT), are discussed to provide context for the datasets used in this study. The chapter then reviews various methods for medical image segmentation, including interactive segmentation and foundation models, while also highlighting the challenges associated with volumetric segmentation. Finally, it introduces continual learning techniques relevant to interactive medical image segmentation.

Chapter 3: Methodology This chapter presents the core methodology of the proposed framework. It begins with a formal definition of the problem, outlining the objectives and constraints of interactive segmentation with continual test-time adaptation. The chapter then details the proposed framework, which consists of a teacher-student learning paradigm to enable real-time adaptation while preserving previously learned knowledge. The teacher model serves as a knowledge repository across slices, while the student model adapts based on user interactions. The key components, including augmentations, consistency loss, and interactive loss, are described in detail. The role of exponential moving average (EMA) in stabilizing the learning process is also discussed.

Chapter 4: Experiments This chapter details the experimental setup and evaluation methodology used to validate the proposed framework. It describes the datasets used for evaluation, including the CHAOS Challenge, LiTS, and AMOS datasets, along with their characteristics. The chapter then presents the details of the segmentation models used in the experiments, such as UNet, UNet++, and DeepLabV3. The experimental results analyze the performance of different models in an interactive segmentation setting, focusing on the impact of test-time adaptation. The chapter also explores the trade-off between consistency loss and interactive loss in achieving optimal segmentation performance.

Chapter 5: Results This chapter presents the results of the proposed interactive segmentation approach. It begins with a quantitative and qualitative evaluation of the iterative interactive segmentation (IIS) method, comparing its performance with conventional interactive segmentation techniques. The results demonstrate the benefits of the continual test-time adaptation framework in improving segmentation accuracy and reducing user interaction efforts. The chapter then compares the proposed method with prompting-based foundation models, presenting a detailed analysis of their relative strengths and weaknesses.

Chapter 6: Conclusion The final chapter summarizes the key findings of the thesis and discusses its contributions to the field of interactive medical image segmentation. It reflects on the improvements achieved through continual test-time adaptation and highlights the broader implications of this work in medical imaging. The chapter concludes with potential directions for future research, including exploring advanced learning strategies for real-time adaptation and integrating the framework with large-scale medical imaging applications.

Overall, this thesis provides a comprehensive investigation into the challenges of interactive segmentation in medical imaging and proposes a novel solution based on continual adaptation at test time. The

presented framework introduces new methodologies that significantly enhance the accuracy, efficiency, and usability of interactive segmentation models, making them more practical for clinical applications.

Chapter 2

Background

2.1 Medical Imaging

Medical imaging encompasses a variety of modalities commonly used in medical image segmentation applications, each differing significantly from natural images. Unlike natural RGB images captured by devices such as phone cameras, medical images are often acquired using specialized equipment designed to visualize anatomical or physiological domains. This chapter provides an overview of some of the most widely used medical imaging modalities, with a primary focus on introducing the imaging modality employed in this work, namely Computed Tomography (CT). In addition, other prominent medical imaging modalities are briefly discussed to highlight the distinctive characteristics of the medical imaging domain compared to the natural image domain.

2.1.1 Radiography (X-Rays)

Radiography, one of the earliest forms of medical imaging, dates back to the 1890s when Wilhelm Röntgen discovered X-rays, a type of electromagnetic radiation. The medical applications of X-rays were rapidly recognized following their discovery [49]. Radiography involves using X-rays to visualize the internal structures of the human body by directing X-ray beams toward a target area, such as an arm, to assess conditions like bone fractures. The absorption of X-rays varies depending on the density and composition of the tissues or materials within the target area. Dense structures such as bones or metallic objects absorb more radiation and appear white on the resulting image, while less dense structures like air or fat absorb less radiation and appear darker [49]. Radiography is categorized into two types: fluoroscopy and projectional radiography [49]. Fluoroscopy provides a continuous real-time visualization of internal body structures, achieved by converting the transmitted X-rays into images on a fluorescent screen. In contrast, projectional radiography generates a single two-dimensional projection of the target area, recorded on photographic film, based on the radiation passing through the body.

2.1.2 Computed Tomography (CT)

Computed Tomography (CT) is a widely used imaging modality for visualizing internal structures within the human body [49]. A CT scan employs a rotating X-ray tube equipped with detectors to capture multiple projections of the target region, such as the abdominal area [50]. As the X-ray beam passes through the target area, the attenuation of the beam is measured, which varies depending on the density of the traversed materials. This differential attenuation enables the distinction of organs, bones, and tissues. The captured projections are processed using computational algorithms to generate 2D slices of the target region, which are subsequently reconstructed into a 3D volume. Each voxel within this volume is assigned a CT number, calculated based on the X-ray attenuation properties of the tissue within that voxel. The CT number is expressed in Hounsfield Units (HU), ranging from approximately -1000 for air, 0 for water, and up to 3000 for dense materials like metal (e.g. artificial implants). CT scans are typically displayed as a 3D stack of 2D grayscale images, where the CT number determines the voxel gray intensity [49, 50].

2.1.3 Magnetic Resonance Imaging (MRI)

Magnetic Resonance Imaging (MRI) leverages strong magnetic fields and radio frequency pulses to generate detailed images of specific target areas, such as the head or abdomen [49, 51], alongside CT. The technique utilizes hydrogen nuclei, which are abundant in the human body, to create images that can be 2D slices that represent the target area in discrete segments or continuous 3D volumes [49]. The underlying principle of MRI relies on the interaction of atomic nuclei with magnetic fields. Atomic nuclei, possessing a spin and a small magnetic moment, align in the direction of a static magnetic field induced by the MRI device. A radio frequency pulse is then applied, disrupting this alignment and forcing the nuclei into a new orientation relative to the magnetic field. When the pulse is discontinued, the nuclei gradually return to their original alignment, releasing electromagnetic energy in the process. The MRI system detects this released energy and distinguishes tissues based on the rate at which the energy is emitted, allowing detailed imaging of anatomical structures [51].

2.2 Medical Image Segmentation

Image segmentation is a fundamental task in computer vision that involves dividing an image into distinct regions based on the desired segmentation objective. It is primarily categorized into semantic segmentation and instance segmentation [52]. Semantic segmentation assigns each pixel or voxel in the image to a specific semantic class from a predefined set of labels, enabling differentiation between object types [52, 53]. For example, in an image captured by a self-driving car, each pixel may be labeled as a pedestrian, car, road, tree, or background, with labels represented by integer values such as 0 for the background and 1, 2, ..., n for respective classes [52]. In contrast, instance segmentation not only identifies target object classes, but also delineates individual instances of each class [52, 53]. For

example, it identifies and outlines all individual pedestrians in the image, providing both localization and instance-specific boundaries [52]. Recently, panoptic segmentation has emerged as a hybrid approach, combining semantic and instance segmentation to assign both a semantic label and an instance identifier to each pixel [52]. This unified method captures both class-level and instance-specific details, broadening the scope of image segmentation.

Medical image segmentation plays a pivotal role in computer-aided diagnosis systems across diverse medical applications. It is a critical step for tasks such as tumor detection, organ delineation, and disease diagnosis. Despite advancements in deep neural networks (DNNs), achieving consistently accurate segmentation remains challenging due to the variability in imaging modalities (e.g., CT, MRI, ultrasound) and the inherent heterogeneity in patient data. Factors such as differences in image quality, noise levels, and patient-specific anatomical variations further exacerbate these challenges, often resulting in segmentation inaccuracies. The domain of medical image segmentation predominantly focuses on tasks of semantic segmentation [54], wherein each pixel in the image is classified into a predefined category, such as the background or the target organ.

2.3 Methods for Medical Image Segmentation

Traditional image segmentation methods are categorized into intensity-based, boundary-based, and region-based approaches. While these methods are relatively straightforward to implement, they often fall short in terms of accuracy and robustness. Intensity-based segmentation relies on pixel intensity values to isolate the region of interest, assuming that the pixels within the target region exhibit similar intensity levels. Thresholding is a commonly employed technique in this category and works effectively when there is a clear contrast between the foreground and background. However, in scenarios where the background is occluded or the boundaries of the region of interest are indistinct, this method encounters significant limitations. One of the key challenges is determining an appropriate threshold value, especially when the intensity values of the background and foreground overlap.

Boundary-based methods focus on detecting edges or boundaries between structures in an image and utilize these as the foundation for segmentation. Edge detection techniques such as the Canny edge detector [55] are widely employed in this category, leveraging gradients to identify edges. However, these methods can be sensitive to noise and may struggle to delineate boundaries in images with complex textures or low contrast.

Region-based segmentation, on the other hand, partitions the image into regions based on pixel similarities. The region growing technique is a notable approach in this category, where segmentation begins with a seed point (a pixel) and iteratively expands by including neighboring pixels that meet predefined inclusion criteria. These criteria often consider intensity or texture similarities. In medical imaging, region growing has shown success in segmenting structures such as tumors and lesions [55]. However, the primary challenges of this method include the selection of an appropriate seed point and defining inclusion criteria, both of which significantly impact the segmentation outcome.

Despite the utility of traditional methods in earlier applications, modern image segmentation techniques predominantly rely on deep learning-based approaches [54–56].Encoder-decoder architectures such as U-Net [18] and its derivative U-Net++ [57] have achieved significant success in medical image segmentation. These architectures have been adapted for 3D volumetric segmentation tasks, with examples including V-Net, a 3D fully convolutional neural network (CNN) proposed by [20], and the 3D extension of U-Net introduced by [19]. Further innovations, such as ConResNet by [58], have incorporated inter-slice context residual learning to enhance performance.

Recent advancements in medical image segmentation have also explored the integration of transformers with CNNs. For instance, UNETR employs a transformer-based encoder to learn sequential representations of volumetric data, capturing global context across multiple scales. Similarly, CoTr introduces an efficient framework for bridging CNNs and transformers, effectively combining their strengths. Despite these advancements, the generalization of medical image segmentation techniques remains an open challenge due to the inherent complexities of medical images. These challenges include low tissue contrast, irregular and highly variable shapes of segmentation targets, diverse imaging and segmentation protocols, and inter-patient variability, all of which complicate the design of robust and generalizable models.

2.3.1 Interactive Medical Image Segmentation

Extensive research has been conducted to investigate interactive image segmentation, employing various interaction modalities such as bounding boxes [29, 30], polygons [31–33], clicks [30, 34–41], and scribbles [28, 42, 43], and combinations of these methods [59]. Recent advancements in click-based interactive segmentation primarily focus on two complementary directions: (1) the design of more effective backbone networks and (2) the development of sophisticated refinement modules built on these backbones.

In the first direction, hierarchical backbone architectures, encompassing both convolutional neural networks (ConvNets) [38, 41] and Vision Transformers (ViTs) [37, 59], have been extensively explored for interactive segmentation tasks. ViTs, in particular, have emerged as a prominent tool in this domain due to their ability to model global dependencies effectively. For instance, FocalClick [36] achieves state-of-the-art segmentation accuracy and computational efficiency by employing SegFormer [45] as its backbone. Similarly, iSegFormer [37] leverages the Swin Transformer [46] as its backbone, offering superior performance in interactive segmentation tasks, particularly within the context of medical imaging.

In the second direction, numerous refinement modules, such as local refinement [36, 38] and click imitation [60], have been introduced to enhance segmentation performance. For instance, DEXTR [44] isolates target objects by leveraging their four extreme points—the left-most, right-most, top-most, and bottom-most pixels—thereby providing a more precise specification of object boundaries. Similarly, FCA-Net [34] highlights the importance of the initial user click in improving segmentation outcomes. Other approaches, such as FocalClick [36] and FocusCut [38], adopt click simulation techniques to

refine segmentation iteratively. These methods introduce local refinement mechanisms that enhance segmentation precision by focusing on erroneous regions. However, click simulation strategies, such as those in FocalClick and FocusCut, often fail to capture the iterative nature of interactive segmentation tasks. In practical applications, users iteratively add clicks in problematic regions based on the current segmentation mask, whereas simulated training clicks are typically generated simultaneously.

2.3.2 Foundation Model for Medical Image Segmentation

Recent advancements in vision foundation models leverage prompting techniques to enable generalization across various tasks. These models, trained on extensive datasets of natural images, demonstrate the ability to segment previously unseen structures based on spatial prompts [61–63] or example-based prompts [64–66]. Notably, some of these foundation models have also been adapted for interactive segmentation through the use of spatial prompts.

However, fine-tuning such models—originally designed for natural image datasets—is often less effective in the biomedical imaging domain compared to training models from scratch [67]. A prominent example of these models, the Segment Anything Model (SAM) [61], exhibits commendable performance on specific biomedical segmentation tasks, particularly those with well-defined boundaries, such as organ segmentation in abdominal CT scans. Nonetheless, its performance significantly diminishes when applied to tasks requiring the delineation of subtle or intricate structures, such as deep brain regions in MRI scans [63, 68, 69].

A comprehensive review by Zhang et al. [70] aggregates findings from multiple studies that assess SAM's zero-shot capabilities across various medical imaging modalities. For instance, SAM achieves performance levels comparable to or exceeding those of established tools like the Brain Extraction Tool (BET) [71] for tasks such as MRI brain extraction and segmentation. However, its efficacy diminishes for more complex tasks, such as brain tumor segmentation from MRI images, where it falls short of state-of-the-art methods [72]. Similarly, studies by Roy et al. [73] indicate that SAM, when applied to organ segmentation in abdominal CT scans, performs competitively using box prompts but struggles to reach state-of-the-art accuracy with point prompts. To address these limitations, Ma et al. [74] introduced Med-SAM, a fine-tuned variant of SAM tailored for medical imaging. Med-SAM was trained on a dataset of approximately 1.5 million image-mask pairs derived from 10 imaging modalities, aiming to create a versatile medical foundation model. This model was evaluated against the zero-shot SAM and task-specific models across 146 segmentation tasks.

The performance of SAM remains task-dependent, excelling in scenarios where the region of interest (ROI) is large and distinct from the background but encountering difficulties with smaller, intricate structures. Ambiguous prompts further exacerbate these challenges. A key limitation of SAM in the medical domain is attributed to the characteristics of the SA-1B dataset, which primarily consists of natural images. These images typically feature sharp object-background boundaries, a stark contrast to the more complex and fragmented nature of biomedical images [61]. Consequently, while foundation models like SAM hold promise, their adaptation to the unique requirements of biomedical imaging re-

mains an open challenge due to the specialized nature of regions of interest and the diversity of imaging modalities in this field.

2.3.3 Issues with Volumetric Interactive Medical Image Segmentation

Volumetric interactive medical image segmentation has proven to be an effective method for achieving precision in medical image analysis. By combining interactive segmentation with foundational models, these methods aim to enhance segmentation accuracy while minimizing user interaction compared to conventional techniques. However, despite these advancements, two major challenges hinder the widespread applicability of these approaches in volumetric interactive medical image segmentation.

The first significant issue arises from the heavy reliance on 2D segmentation methods in many existing approaches [75, 76], which fail to account for the inherently volumetric nature of 3D medical imaging data. In these methods, segmentation is performed on a slice-by-slice basis, treating each 2D image in isolation and disregarding the spatial continuity and volumetric priors available in 3D data. This independent treatment of slices leads to a substantial increase in the user annotation burden, as each slice must be segmented individually without leveraging the correlations between adjacent slices. Although 3D networks [11, 19, 20] have been developed to address this issue by utilizing higher-order volumetric features, they come with significant computational demands. The high parameter count and increased memory requirements of 3D models often necessitate trade-offs, such as reduced resolution or simplified architectures, to fit within the constraints of available hardware. These compromises can degrade the segmentation accuracy and limit the usability of these models, particularly in applications involving large and complex volumetric datasets.

The second critical challenge lies in the lack of flexibility of existing methods for human-in-the-loop workflows, which are essential in interactive segmentation tasks. Most current approaches require users to manually inspect individual slices for segmentation inaccuracies and provide corrective inputs to refine the segmentation output. This manual intervention is time-consuming and inefficient. Additionally, traditional interactive image segmentation (IIS) techniques and prompting foundation models further compound this inefficiency by requiring annotators to input corrections repeatedly for each slice. These methods do not capitalize on the continuity and similarities between adjacent slices, as they lack the capability to retain contextual information across slices. Consequently, the absence of inter-slice contextual retention leads to suboptimal segmentation performance and increases the overall time and effort required for annotation.

Addressing these challenges is critical for improving the effectiveness and efficiency of volumetric interactive medical image segmentation. By leveraging inter-slice spatial relationships, advanced segmentation methods can retain contextual knowledge across slices, reducing the need for repetitive user inputs and enhancing the accuracy and robustness of the segmentation process. Such advancements are essential for making volumetric segmentation workflows more practical and scalable for clinical and research applications.

2.4 Continual Learning (CL) for Volume Segmentation

In practical applications, iterative corrections through user-driven prompts are often essential to refine outcomes until they meet specific criteria [41]. However, research into human-in-the-loop approaches for medical interactive segmentation remains limited [62, 63], and existing methods have demonstrated suboptimal performance. One significant challenge is the efficiency of these models in terms of training and user interaction. While training models from scratch on extensive datasets [40,41,62,63,74] can enhance their representational power, this process is resource intensive, time consuming, and computationally demanding. In addition, domain gaps inherent in medical imaging often lead to models that lack specificity for particular applications, necessitating increased user interaction to achieve satisfactory results [62].

Deep learning models, including those designed for medical image segmentation, are generally trained on static datasets that assume a fixed data distribution representative of future scenarios. However, in real-world medical applications, models often encounter data distributions that differ significantly from the training set, leading to degraded performance. To address this challenge, various domain adaptation techniques have been proposed, including unsupervised domain adaptation [77–80], test-time training [81], test-time adaptation [82], and continual learning [83, 84]. Although some of these methods require access to target data during training or subsets of source data for adaptation, they face significant practical constraints. In most real-world medical imaging scenarios, the target domain is unknown during training, and retaining source data is often unfeasible due to ethical, legal, or computational constraints. Additionally, target domains in medical imaging are rarely static, as data distribution may evolve over time due to variations in imaging protocols, devices, and patient populations.

The reliance on static datasets and predefined segmentation tasks underscores a significant limitation in conventional paradigms for medical image segmentation. Models developed and trained within closed, static environments are often ill-equipped to handle the dynamic and evolving nature of real-world medical imaging, where data distributions frequently shift and tasks continually change. These limitations highlight the pressing need for methodologies that enable continual test-time adaptation in medical image segmentation. This approach addresses the inherently fragmented and evolving landscape of medical imaging by allowing models to dynamically adjust to changing target domains without requiring access to the original training data. Such adaptability ensures consistent and reliable performance across diverse and shifting medical imaging scenarios, ultimately enhancing both clinical applicability and patient care.

Existing methods also exhibit significant inflexibility in supporting human-in-the-loop segmentation workflows. Typically, these approaches require users to manually inspect mis-segmented slices and provide iterative corrections for refinement, resulting in labor-intensive and inefficient processes. Furthermore, traditional interactive image segmentation (IIS) techniques and prompting foundation models exacerbate these challenges, as annotators must provide inputs for each slice individually. This repetition is not only inefficient but also fails to leverage the inherent continuity and similarities between

adjacent slices, as these models lack mechanisms to retain or utilize contextual knowledge beyond the current slice.

To address these shortcomings, we propose an innovative framework that integrates principles from continual learning into interactive segmentation within the context of a continual test-time adaptation framework [84]. Our approach leverages the observation that segmentation masks between consecutive slices in medical imaging, such as CT or MRI volumes, tend to exhibit significant overlap. By using the user-provided corrections for refining the segmentation mask of a given slice, our framework employs continual adaptation methods to update the model parameters dynamically. This enables the model to adapt not only for the current slice but also for subsequent slices and volumes, thereby improving segmentation efficiency and accuracy across the entire imaging dataset.

Chapter 3

Methodology

In this section, we examine the methods and algorithms utilized to achieve the primary objective of this research: enhancing the segmentation performance of a pre-trained model on previously unseen medical image volumes while ensuring that its performance on prior data remains unaffected. The study is structured to address the dual challenge of improving segmentation accuracy for novel datasets while maintaining generalizability to prevent overfitting. This includes a detailed discussion of the iterative teacher-student framework adopted for interactive segmentation. In this architecture, the student model is designed to adapt rapidly to new domains, while the teacher model incrementally incorporates the newly acquired knowledge without succumbing to catastrophic forgetting. We also describe our method configurations and highlight the challenges encountered during implementation. The solutions devised to address these challenges are detailed, emphasizing their significance in achieving the project's objectives.

Building on this foundation, we elaborate on the rationale behind the dual objective, focusing on the strategies implemented to preserve the model's performance on previously trained data. This serves as a safeguard against overfitting to the current slice, which could otherwise compromise accuracy and increase annotation efforts in subsequent slices during interactive segmentation. Finally, we outline the test plans designed to evaluate the proposed approach, detailing the metrics used to measure its efficiency and accuracy in reducing the annotation burden for future slices while ensuring robust segmentation performance.

3.1 Problem Definition

Let t denote a segmentation task consisting of image-segmentation pairs $\{(x_j^t, y_j^t)\}_{j=1}^N$. In step i, given an image x^t, a set of user interactions u_i, and the previous prediction $\hat{y}_{i-1}^t$, we define a function $f_\theta(x^t, u_i, \hat{y}_{i-1}^t)$ parameterized by θ, which generates a segmentation result $\hat{y}_i^t$. The set of user interactions u_i - which may consist of positive or negative scribbles - is provided by a user who has access to the image x^t and the prior prediction $\hat{y}_{i-1}^t$.

The continuous adaptation of a pre-trained interactive segmentation model can be formally expressed as follows. Let $f_{\theta_0}(x)$ represent the pre-trained segmentation model, where θ_0 are the parameters

learned from the source training dataset D_s. Due to privacy concerns, source training data D_s is typically unavailable during the testing phase. During this phase, the model encounters a new set of test data D_t, which has not been previously observed or annotated.

In this framework, the student model leverages user-provided interactions u_i - including positive or negative scribbles - to refine its predictions. These interactions, combined with the pseudo-segmentation mask generated by the teacher model, are utilized in a test-time training (TTT) approach. The student model adapts iteratively, updating its parameters based on the user-annotated scribbles and the guidance from the teacher model's pseudo-labels. This adaptation process ensures that the student model remains flexible and improves segmentation accuracy for the test data, even in the absence of the original source training dataset. By integrating this teacher-student paradigm, the approach mitigates catastrophic forgetting while effectively incorporating new knowledge during the testing phase.

3.2 Framework

Our proposed framework employs an iterative approach that integrates a teacher-student architecture with test-time training, enabling on-the-fly adaptation of any pre-trained, off-the-shelf model without requiring modifications to its architecture. Both the teacher and student models share the same architecture and initial weights. The framework utilizes test-time augmentations and a low-variance thresholding technique to generate an averaged pseudo-segmentation mask for each slice. The student model undergoes test-time training, leveraging user-provided scribbles and the pseudo-segmentation mask for updates. In parallel, the teacher model is updated using an exponential moving average (EMA) of the student model's parameters.

Unlike prior methods [81, 82, 84], which reset the model's weights after processing each slice, effectively discarding accumulated knowledge, our approach ensures that the teacher model functions as a repository for knowledge across slices. The student model, in turn, is optimized for rapid adaptation to the current slice. This dynamic interaction strikes a balance between retaining knowledge from previous slices and incorporating information from the current slice. Notably, a simplistic strategy of continually updating the student model across slices without the guidance of a teacher model is suboptimal, as test-time adaptation (TTA) can result in overfitting to the current slice, thereby degrading performance on subsequent slices. The inclusion of the teacher model serves as a critical regularization mechanism, preserving knowledge across slices without succumbing to overfitting.

In the subsequent subsections, we provide an in-depth discussion of the individual components of the framework. This includes a detailed explanation of the teacher model and the necessity of test-time augmentations, the role of the student model in incorporating user-scribbles and updating its parameters, and the adaptation of the teacher model through exponential moving averages.

3.2.1 Teacher Model

In this study, we employ a pre-trained segmentation model, parameterized by θ_0, which processes a slice from a medical volume, denoted as x_t, as its input. Both the teacher model $(f_{\theta^T}(x_t))$ and the student model $(f_{\theta^S}(x_t))$ are initialized using the same pre-trained interactive segmentation weights, ensuring consistency at the outset, such that $\theta^T = \theta_0$ and $\theta^S = \theta_0$. The initialization ensures that both models begin with a shared representation learned from the pre-trained weights, which forms the foundation for subsequent adaptations and refinements.

The teacher model, $(f_{\theta^T}(x_t))$, is leveraged to generate multiple predictions $\hat{y}_i^T$ by applying test-time augmentations (TTA) to the input data, as described in prior works [48, 84]. These augmentations are exclusively intensity-based transformations, including operations such as the addition of Gaussian noise, blurring, and pixel intensity inversion. These augmentations aim to increase the robustness of the predictions by simulating various input conditions that the model might encounter during deployment.

However, while traditional TTA methods improve prediction reliability by introducing variability during inference, they are inherently limited in addressing domain shifts between the source and target data distributions. This limitation has been documented in prior studies [81, 82, 84], highlighting that such methods lack the capacity to adapt to the unique characteristics of the target domain. Consequently, while the teacher model provides augmented predictions, it is not inherently capable of aligning with significant differences in data distribution, necessitating further innovations to bridge this gap.

3.2.1.1 Augmentations

To address this challenge, we propose a method that dynamically adjusts the number of augmentations based on the teacher model's confidence levels. The confidence of the teacher model is quantified by applying the softmax function to its output masks, as suggested in prior work [85]. From these output masks, we extract pixel-wise confidence values. Specifically, we compute the ratio of the number of pixels with prediction confidences falling within the range of 0.4 to 0.6 to the total number of pixels in the image. This ratio, denoted as *ratio*$(f_{\theta^T}(x_t))$, reflects the uncertainty in the prediction of the model, as confidence values in the range of 0.4 to 0.6 signify ambiguous predictions where the model is uncertain to assign a pixel to a specific class. One challenge in using the ratio of pixels with prediction confidences falling within the range of 0.4 to 0.6 to the total number of pixels lies in the disparity between the number of uncertain pixels and the total number of pixels in the medical slice. To address this issue, we introduce a normalized ratio that adjusts the influence of uncertain pixels. Specifically, the ratio is defined as:

$$\textit{ratio}(f_{\theta^T}(x_t)) = \frac{\textit{uncertain pixels}(0.4 \leq \textit{pred. conf.} \leq 0.6)}{\textit{total pixels in image}} \times \frac{\textit{total pixels}}{\textit{total uncertain pixels}} \tag{3.1}$$

Here, *total uncertain pixels* refers to the total number of pixels with prediction confidences falling within the specified range (0.4 to 0.6). This ratio 3.1 accounts for the sparse nature of uncertain pixels in

large images and ensures their proportional representation in the ratio calculation. This ratio regulates the number of augmentations. Equation shows the number of augmentations used based on this ratio.

$$y'_{t^T} = \frac{1}{N}\sum_{j=1}^{N} f_{\theta^T}(\tilde{x}_t^j) \begin{cases} N = 4 & \text{for} \quad \text{ratio}(f_{\theta^T}(x_t)) \leq 0.3 \\ N = 16 & \text{for} \quad 0.3 < \text{ratio}(f_{\theta^T}(x_t)) \leq 0.7 \\ N = 32 & \text{for} \quad 0.7 < \text{ratio}(f_{\theta^T}(x_t)) < 1 \end{cases} \tag{3.2}$$

In this equation, $f_{\theta^T}(x_t)$ represents the teacher model's output mask for the input x_t, and N denotes the number of augmentations applied. When the ratio of uncertain pixels ($0.4 \leq$ confidence ≤ 0.6) is low (ratio ≤ 0.3), a smaller number of augmentations ($N = 4$) is applied, as the model is relatively confident. Conversely, when the ratio is high ($0.7 <$ ratio < 1), the number of augmentations increases ($N = 32$) to compensate for the model's uncertainty. For moderate uncertainty levels ($0.3 <$ ratio ≤ 0.7), an intermediate number of augmentations ($N = 16$) is utilized.

3.2.1.2 Consistency Loss

In this context, y'_{t^T} denotes the mean segmentation mask derived from the teacher model, while $f_{\theta^T}(\tilde{x}_t^j)$ represents the segmentation output for an augmented version of the target slice x_t. The parameter N indicates the number of augmentations applied, and $f_{\theta^T}(x_t)$ corresponds to the prediction confidence of the pre-trained teacher model for the current input $\tilde{x}_t$. To mitigate overconfidence in the predicted segmentation masks, temperature scaling [86] with a parameter $\tau > 1$ is employed. This process generates a smoothed segmentation mask by averaging multiple predictions from the teacher model across various augmentations, denoted as $\hat{y}_T^a$.

For each slice x_t, the student model produces an initial segmentation mask $\hat{y}^S = f_{\theta^S}(x_t)$. Subsequently, the cross-entropy loss is computed at the pixel level by comparing the segmentation mask from the student model with the mean segmentation mask generated by the teacher model. This can be expressed as follows:

$$\mathcal{L}_{consistency} = -\text{sum}\left(\sum_{c} \hat{y}_T^a[:, c] \log\left(\hat{y}_S[:, c]\right)\right). \tag{3.3}$$

Here, we focus only on the pixels within the segmentation mask that exhibit low variance across the augmented versions. This prioritizes regions with consistent predictions across different augmentations, leading to a more reliable pseudo-segmentation mask.

3.2.2 Student Model — User Interaction

The integration of user feedback into the segmentation process is facilitated through a scribble mask, m, and a label vector, $\hat{y}_U$. The scribble mask, m, is an n-dimensional binary vector where each element equals 1 if user feedback has been provided for the corresponding pixel (via a scribble) and 0 otherwise. Correspondingly, the label matrix, $\hat{y}_U$, is represented as an $n \times c$ matrix, with each row comprising a

one-hot encoded vector specifying the user-provided label for a given pixel, while the remaining entries are zero. To emulate thicker user-provided scribbles, the scribble mask m undergoes a blurring process, following which the user-defined label is assigned to all non-zero pixels within the blurred mask.

For cases involving multiple scribbles, this process is performed individually for each scribble, after which the resultant masks and label matrices are combined through an element-wise addition. In instances of overlapping scribbles, no new information is incorporated for the overlapping pixels. For ease of representation, and with a slight abuse of notation, the combined scribble mask and label matrix are still denoted as m and $\hat{y}_U$, respectively.

3.2.2.1 Interactive Loss

When processing multiple scribbles, the steps outlined above are repeated for each scribble individually. The resulting masks and label matrices are then merged using element-wise addition. If there are overlapping regions between the masks, the overlapping pixels are excluded from contributing new information. To maintain consistency, the merged mask and label matrix continue to be represented as m and $\hat{y}_U$, respectively.

The interactive loss, $\mathcal{L}_{interactive}$, is formulated as:

$$\mathcal{L}_{interactive} = -\text{sum}\left(\sum_{c} m[:]\ \hat{y}_U[:,c] \log\left(\hat{y}_S[:,c]\right)\right), \tag{3.4}$$

where $m[:]$ refers to the binary mask indicating the region influenced by the user-provided scribbles, $\hat{y}_U[:,c]$ denotes the one-hot encoded ground truth corresponding to class c, and $\hat{y}_S[:,c]$ represents the predicted probability for class c generated by the model. This loss function serves to penalize discrepancies between the model's predicted probabilities and the user-defined ground truth within the scribbled regions, enabling the model to iteratively refine its segmentation output based on the feedback provided by the user.

We define the cumulative loss, $\mathcal{L}_t$, as the sum of individual loss terms $\mathcal{L}_{consistency}$ and $\mathcal{L}_{interactive}$. We update student model parameters from $\theta_t^S \rightarrow \theta_{t+1}^S$ by performing standard stochastic gradient descent on $\mathcal{L}_t$.

3.2.3 Exponential Moving Average

After updating the student model, the teacher model is subsequently refined to ensure that it retains the essential knowledge from prior training iterations while incrementally integrating new information acquired from the student's adaptation. This transfer of knowledge is accomplished by employing an Exponential Moving Average (EMA) to update the teacher model's parameters based on the student's parameters:

$$\theta_{t+1}^T = \alpha\, \theta_t^T + (1-\alpha)\, \theta_t^S, \tag{3.5}$$

where α serves as a smoothing factor that regulates the contribution of the teacher's past parameters and the student's current parameters. The final prediction for a given slice x is obtained as the segmentation mask generated by the updated teacher model.

The EMA mechanism plays a pivotal role in addressing the issue of catastrophic forgetting, a significant challenge in scenarios involving continual learning. By progressively integrating new knowledge while preserving critical insights from prior iterations, this approach ensures robust and stable performance in interactive segmentation tasks.

Chapter 4

Experiments

To begin, we outline the problem definition and provide a comprehensive description of the four datasets employed in this study: the CHAOS challenge dataset, the LiTS dataset, the AMOS challenge dataset, and the DSB dataset. These datasets form the basis for the subsequent analyses and experiments conducted in this work. Next, we delve into the model configurations essential for this research.

4.1 Datasets

In this section, the details of the three datasets that we used for analysis during this project are explained. These include the CHAOS challenge dataset, the LiTS dataset, the AMOS challenge dataset, and the DSB dataset.

4.1.1 CHAOS Challenge Dataset

The CHAOS (Combined Healthy and Abnormal Organ Segmentation) Challenge serves as a prominent benchmark for assessing and enhancing algorithmic performance in segmenting both normal and diseased organs within medical imaging datasets [87]. In this study, a subset of the CHAOS dataset was employed, comprising computed tomography (CT) images of 40 individuals identified as potential liver donors. All subjects included in this dataset exhibited no indications of liver abnormalities such as tumors or other pathological conditions. The CT imaging primarily focuses on the upper abdominal region and was conducted during the portal venous phase, a period approximately 70–80 seconds after contrast agent injection or 50–60 seconds post-bolus tracking. This acquisition phase optimally highlights liver parenchyma by leveraging the blood supply from the portal vein.

Three distinct CT imaging devices were utilized in generating the dataset: Philips SecuraCT with 16 detectors, Philips Mx8000 CT with 64 detectors, and Toshiba AquilionOne with 320 detectors, all of which were equipped with spiral CT functionality. Uniform protocols were adhered to in subject positioning and image alignment to ensure consistency across datasets. The images are provided in 16-bit DICOM format, featuring a spatial resolution of 512×512, with x-y pixel spacing varying between 0.7 mm and 0.8 mm and an inter-slice distance (ISD) ranging from 3.0 mm to 3.2 mm. Each volumetric scan contains an average of 90 slices, with a range spanning from 77 to 105 slices. This results in

a comprehensive dataset of 1367 slices designated for training and 1408 slices allocated for testing purposes.

4.1.2 LiTS Dataset

The ISBI 2017 Liver Tumor Segmentation Challenge (LiTS) dataset [88], a key resource for this study, consists of 200 contrast-enhanced abdominal CT scans in Nifti format, collected from six medical centers. Of these, 130 scans are designated for training, while 70 are reserved for testing. Each scan is represented as a three-dimensional volume and is accompanied by a corresponding ground truth segmentation file. The scans have a fixed resolution of 512×512 pixels in the (x, y) plane, while the number of slices along the z-axis varies significantly, ranging from as few as 75 layers to over 800 layers in some cases. The dataset has been a cornerstone of the LiTS benchmark challenge, which aims to stimulate the development of cutting-edge deep learning models for liver and liver tumor segmentation. The LiTS benchmark challenge addresses critical clinical needs, such as improved diagnostic accuracy and surgical planning. It also facilitates collaborative research efforts and drives technological advancements in the field. The challenge ensures that the models developed are not only effective but also robust, making them adaptable to the diverse and variable conditions encountered in real-world clinical environments.

4.1.3 AMOS Dataset

The AMOS dataset comprises both CT and MRI imaging data obtained from anonymized patients across various clinical medical centers [89]. This dataset includes comprehensive segmentation annotations, featuring 500 CT scans and 100 MRI scans collected from multi-center, multi-vendor, multi-modality, multi-phase, and multi-disease patients. Each scan is annotated at the voxel level for 15 abdominal organs, including the spleen, right and left kidneys, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right and left adrenal glands, duodenum, bladder, and prostate/uterus. This extensive dataset offers challenging examples, serving as a valuable testbed for developing and evaluating robust segmentation algorithms in diverse scenarios and target structures.

To maximize its utility for both the research community and clinical applications, the dataset is distributed in multiple formats, such as DICOM and NIFTI, ensuring compatibility with various workflows and platforms. However, the dataset does not include detailed information regarding sequence types or patient demographics, which may limit certain analyses or interpretations. Despite this limitation, the AMOS dataset remains a critical resource for advancing segmentation techniques under diverse and complex conditions.

4.2 Model Details

In this section, we demonstrate that our proposed method is agnostic to the choice of backbone architecture, thereby enabling its integration with a variety of pre-trained models. To validate this claim,

we performed experiments utilizing different backbone architectures, including UNet, UNet++, and DeepLabv3. These models were chosen for their widespread application and proven effectiveness in medical image segmentation tasks. By showcasing consistent performance improvements across these diverse backbones, we highlight the adaptability and robustness of our framework to operate with any arbitrary model architecture. This flexibility underscores the generalizability of our approach, making it suitable for a wide range of segmentation tasks in medical imaging.

4.2.1 UNet

The U-Net architecture is a convolutional neural network (CNN) specifically designed for biomedical image segmentation. It consists of a contracting path to extract features and reduce spatial resolution and an expansive path to upsample and reconstruct the input's spatial dimensions. A key feature of U-Net is the use of skip connections, which preserve fine-grained spatial details by linking the contracting and expansive paths, enabling precise segmentation.

In our method, U-Net serves as the backbone due to its proven effectiveness in medical image segmentation. Its ability to provide baseline results and retain critical spatial details makes it an ideal choice for evaluating the performance of our approach in comparison to other frameworks.

4.2.2 UNet++

U-Net++ is an advanced semantic segmentation architecture based on U-Net, designed to address limitations in traditional feature fusion. It introduces nested dense skip connections to bridge the semantic gap between encoder and decoder feature maps, enhancing gradient flow and improving segmentation performance. Additionally, deep supervision acts as a regularization mechanism during training, further boosting model accuracy. These innovations make U-Net++ a robust choice for tasks requiring precise feature refinement and segmentation.

4.2.3 DeepLab V3

DeepLabv3 is a semantic segmentation architecture designed for capturing multi-scale contextual information. It leverages Atrous Spatial Pyramid Pooling (ASPP), which applies atrous convolutions with varying dilation rates to extract features at multiple scales effectively. Additionally, it incorporates depth-wise separable convolutions to improve computational efficiency. The model provides accurate segmentation by refining predictions through bilinear interpolation and optional post-processing with a fully connected Conditional Random Field (CRF). This combination makes DeepLabv3 highly effective for complex segmentation tasks.

4.3 Experimental results

In our implementation, we initialized the U-Net, U-Net++, and DeepLabv3 architectures with specific hyperparameters for optimal performance. The configurations included setting encoder depth=4, encoder weights='imagenet', and defining decoder channels=(256, 128, 64, 32) to structure the decoder. These settings ensured a consistent and comparable framework across the various backbone models employed in our experiments.

4.3.1 Performance Comparison of Models

The evaluation of deep learning models for medical image segmentation often necessitates a comprehensive comparison across multiple dimensions, including accuracy and adaptability to iterative feedback. In this study, we focus on three widely used architectures — U-Net, UNet++, and DeepLabV3 — analyzing their performance on diverse datasets, namely CHAOS, LiTS, and CHAOS (MRI). The analysis encompasses both the initial baseline Dice score and the final Dice score achieved after incorporating feedback through interactive iterations. Additionally, the computational aspects, including the average number of iterations required to achieve the final score and the total time spent in these iterations, are evaluated to provide a holistic view of model efficiency. This subsection presents a detailed comparison of the models, highlighting their strengths and limitations in the context of medical image segmentation tasks.

Dataset	Model	Initial Dice Score	Final Dice Score	Avg. No. of Iterations to get the Final Dice Score	Total Time in Iterations
CHAOS	UNet	0.643	0.936	3	35 sec
	UNet++	0.742	0.885	4	40 sec
	DeepLabV3	0.783	0.891	4	57 sec
LiTS	UNet	0.673	0.924	2	30 sec
	UNet++	0.701	0.872	4	42 sec
	DeepLabV3	0.762	0.898	4	50 sec
CHAOS (MRI)	UNet	0.615	0.931	3	53 sec
	UNet++	0.726	0.846	4	68 sec
	DeepLabV3	0.752	0.895	5	72 sec

Table 4.1 Performance comparison of models across datasets. Note: The reported results are specific to slices containing the organ.

The comparative analysis of U-Net, UNet++, and DeepLabV3 highlights significant differences in their performance, adaptability, and computational efficiency in medical image segmentation tasks. Despite U-Net starting with a lower baseline Dice score compared to UNet++ and DeepLabV3, it demonstrated remarkable adaptability during interactive learning. After three interactions on each slice, U-Net achieved the highest final Dice score, underscoring its ability to effectively leverage feedback for per-

formance improvement. This rapid improvement suggests that U-Net benefits from an efficient gradient flow and faster weight updates, allowing it to converge more quickly compared to its counterparts.

In contrast, UNet++ and DeepLabV3 exhibited superior initial performance, likely due to their more complex architectures, which enhance their ability to extract semantic features and capture spatial details. However, the incremental improvement in their performance through interactions was less pronounced, indicating a reduced sensitivity to iterative feedback compared to U-Net. Additionally, the higher computational demands of these models further highlight the efficiency of U-Net, particularly in scenarios with constrained resources or strict time requirements.

Overall, these findings suggest that U-Net, with its simplicity and adaptability, provides an effective baseline for interactive segmentation tasks. While UNet++ and DeepLabV3 may be better suited for applications requiring high initial accuracy, U-Net excels in iterative scenarios, achieving superior performance with fewer iterations and lower computational cost. This adaptability makes it an ideal choice for scenarios involving dynamic feedback or resource limitations, underscoring its utility in practical medical imaging applications.

4.3.2 Balancing Consistency Loss and Interactive Loss

In medical image segmentation tasks, optimizing loss functions is a critical factor for improving model performance. In our experiments, we focused on two distinct loss functions: consistency loss, which ensures alignment between predictions across different resolutions, and interactive loss, which incorporates user feedback to refine the segmentation. During initial trials, we observed that the consistency loss was dominating the interactive loss, limiting the model's ability to effectively utilize user feedback.

To address this issue, we conducted a series of experiments to balance the two losses. Initially, the losses were combined without any scaling, leading to suboptimal updates primarily driven by consistency loss. This was evident in the minimal impact of user feedback on segmentation results, as the consistency loss overshadowed the contribution of the interactive loss. To counter this, we introduced a scaling factor β to amplify the interactive loss, thus balancing the cumulative loss function. The updated loss function was defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{consistency}} + \beta \cdot \mathcal{L}_{\text{interactive}}, \tag{4.1}$$

where $\beta > 1$ was empirically adjusted to achieve the desired balance. After testing various values of β, we found that setting $\beta = 2$ provided the best trade-off between consistency and interaction-driven refinements.

This adjustment resulted in noticeable improvements. The segmentation accuracy, as measured by the Dice score, increased significantly after incorporating user feedback. Qualitative analysis revealed that user-provided corrections were better reflected in the segmentation maps, demonstrating that the model became more responsive to interactive inputs. However, a closer examination of the results high-

lighted an important observation: in certain cases, such as those represented in Segmentation Mask 2 of Figure 4.1, the model failed to incorporate the user interactions effectively. Despite explicit corrective inputs, the segmentation mask did not show any significant updates, indicating that the model was inadequately weighting the interactive loss during optimization. This behavior was attributed to the dominance of the consistency loss, which overshadowed the contribution of the interactive loss.

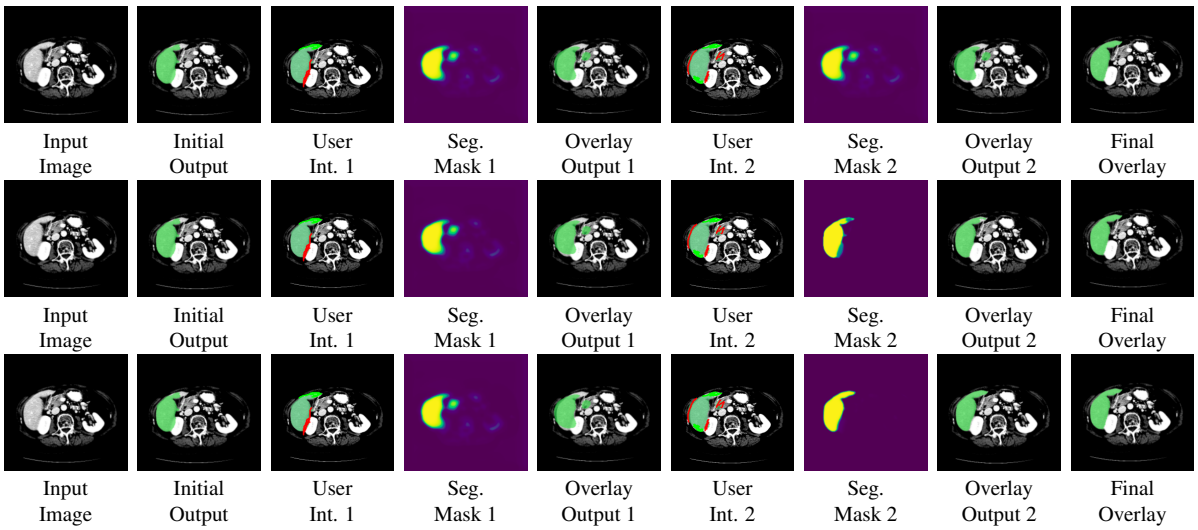

Figure 4.1 Comparison of segmentation outputs across different user interaction levels.

To address this limitation, the introduction of a scaling factor proved instrumental. By amplifying the interactive loss, the model was able to better integrate user-provided feedback. The impact of this adjustment can be seen in the subsequent rows of Figure 4.1, where the segmentation masks exhibit clear improvements following user interactions. These refined masks show enhanced alignment with user corrections, highlighting the effectiveness of the scaling strategy. Additionally, this modification ensured that both loss components, consistency and interactive, were optimally balanced, allowing the model to leverage the strengths of both.

We test $\beta \in \{1.0, 1.5, 2.0, 2.5, 3.0\}$ in Eq. 4.1.

β	Time (min)	Final Dice	Stability
1.0	18.6 ± 0.9	0.89 ± 0.018	User ignored
1.5	11.2 ± 0.6	0.91 ± 0.016	Suboptimal
2.0	$\mathbf{9.24 \pm 0.4}$	$\mathbf{0.936 \pm 0.012}$	**✓Optimal**
2.5	10.1 ± 0.5	0.91 ± 0.015	Unstable
3.0	12.4 ± 0.7	0.88 ± 0.019	Divergent

Table 4.2 Loss Scaling Sensitivity (CHAOS-CT, 5 runs).

Figure 4.1 reveals an important optimization challenge in interactive medical image segmentation. When consistency-based supervision and user-driven interactive supervision are combined with equal weighting, the optimization process becomes dominated by the consistency loss. As a result, corrective

user inputs may have a limited influence on the updated segmentation, even when explicit errors are indicated through scribbles. This behavior is undesirable in interactive settings, where rapid incorporation of user intent is essential.

To explicitly regulate the relative contribution of user feedback, we introduce a scaling factor β on the interactive loss, leading to the combined objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{consistency}} + \beta \cdot \mathcal{L}_{\text{interactive}}. \tag{4.2}$$

We evaluate $\beta \in \{1.0, 1.5, 2.0, 2.5, 3.0\}$ and analyze its impact on segmentation accuracy, convergence speed, and training stability. Quantitative results are summarized in Table 4.3.2. When $\beta = 1.0$, the model largely ignores user corrections, as consistency supervision overwhelms sparse interactive signals. Increasing β improves responsiveness to user input, with $\beta = 2.0$ consistently yielding the best trade-off between stability and adaptability. Larger values ($\beta \geq 2.5$) introduce optimization instability and degrade final Dice performance.

Importantly, the optimal value $\beta = 2.0$ generalizes across datasets without re-tuning. This choice compensates for the inherent imbalance between dense pixel-level consistency supervision and sparse user-provided annotations, allowing both objectives to contribute meaningfully during optimization.

Qualitative examples in Fig. 4.1 corroborate these findings. With appropriate loss scaling, user interactions are faithfully reflected in the updated segmentation masks, leading to progressive and visually coherent refinements across iterations.

4.3.2.1 Component Analysis: Role of Consistency Loss (L_1)

We further isolate the role of consistency loss by comparing three configurations: (i) a manual *copy-previous-slice* baseline, (ii) an interactive-only model trained without consistency regularization or EMA updates, and (iii) the full proposed framework. Experiments are conducted on CHAOS-CT, LiTS, and AMOS datasets using a U-Net backbone.

Configuration	**CHAOS-CT**	**LiTS**	**AMOS**
Copy Previous (Manual)	32.0 ± 1.2	39.0 ± 1.8	79.0 ± 3.2
Interactive Only (w/o L_1)	15.0 ± 0.8	22.0 ± 1.1	40.0 ± 2.1
Full Method (w/ L_1)	$\mathbf{9.24 \pm 0.4}$	$\mathbf{14.0 \pm 0.6}$	$\mathbf{28.0 \pm 1.3}$

Table 4.3 Consistency Loss Contribution. User time (minutes) to Dice 0.95 (5 runs $\pm$ 95% CI).

Results in Table 4.3.2.1 demonstrate that consistency regularization yields a $1.6\times$ to $2.8\times$ reduction in user annotation time to reach a target Dice score of 0.95. Without consistency supervision, user scribbles introduce localized corrections but fail to propagate reliably across slices, leading to unstable predictions and increased interaction effort. In contrast, consistency loss enables robust knowledge transfer across slices by enforcing agreement with temporally aggregated pseudo-labels, significantly improving efficiency and stability.

4.3.2.2 Test-Time Augmentation Effectiveness

The quality of teacher-generated pseudo-labels plays a crucial role in consistency-based learning. We therefore evaluate the impact of adaptive test-time augmentation (TTA) on final segmentation performance. As shown in Table 4.3.2.2, adaptive augmentation with a maximum of $N \leq 16$ transformations consistently outperforms both no-augmentation and fixed-augmentation baselines across all datasets.

Strategy	**CHAOS-CT**	**LiTS**	**AMOS**
No Augmentation	0.781 ± 0.023	0.741 ± 0.028	0.761 ± 0.025
Fixed $N = 6$	0.862 ± 0.019	0.852 ± 0.021	0.832 ± 0.022
Adaptive ($N \leq 16$)	$\mathbf{0.932 \pm 0.012}$	$\mathbf{0.922 \pm 0.014}$	$\mathbf{0.912 \pm 0.015}$
Adaptive ($N \leq 32$)	0.917 ± 0.016	0.901 ± 0.018	0.896 ± 0.019

Table 4.4 Adaptive TTA Impact. Final Dice after fixed interactions (5 runs $\pm$ 95% CI).

Increasing the number of augmentations beyond this threshold introduces diminishing returns and, in some cases, performance degradation due to augmentation-induced artifacts. These results validate the use of uncertainty-driven adaptive augmentation, which balances pseudo-label reliability with computational efficiency.

4.3.2.3 EMA Momentum Sensitivity (α)

EMA momentum α (Eq. 3.5) balances retention vs. adaptation. We test $\alpha \in \{0.70, 0.80, 0.90, 0.95, 0.99, 0.995\}$ on CHAOS-CT.

α	Final Dice	Interactions	Behavior
0.70	0.840 ± 0.021	20 ± 2	Rapid forgetting
0.80	0.870 ± 0.019	15 ± 1	Improving
0.90	0.890 ± 0.017	12 ± 1	Stable
0.95	0.901 ± 0.015	8 ± 1	Very good
0.99	$\mathbf{0.932 \pm 0.012}$	$\mathbf{4 \pm 1}$	**✓Optimal**
0.995	0.910 ± 0.016	9 ± 1	Rigidifies

Table 4.5 EMA Momentum Optimization (CHAOS-CT, 5 runs).

We analyze the sensitivity of the exponential moving average (EMA) momentum parameter α, which controls the balance between adaptation to the current volume and retention of prior knowledge. Results in Table 4.3.2.3 indicate that $\alpha = 0.99$ achieves the best overall performance, maintaining stable segmentation while enabling effective adaptation across slices.

Chapter 5

Results

5.1 Evaluation of Interactive Segmentation Approach

We conducted a comprehensive evaluation of our interactive segmentation approach against other state-of-the-art methods across four medical imaging datasets: CHAOS (CT and MRI Slice Data), LiTS, AMOS, and the 2018 Data Science Bowl (microscopic cell data). These datasets consist primarily of CT volume data, with CHAOS also containing MRI data. The evaluation included the 95% confidence interval (CI) of segmentation performance obtained using manually provided scribbles and clicks over three interaction loops across different methods. Each interaction loop could involve multiple user-provided scribbles or clicks.

5.1.1 Quantitative Results with Interactive Iterative Segmentation (IIS)

Table 5.1 underscores the significant efficiency gains achieved by our framework in reducing user annotation time across all evaluated datasets, as validated by a medical expert. Specifically, when annotations were performed entirely from scratch without external computational aid (first row), our method demonstrated substantial improvements, reducing user annotation time by factors of 4.60 (CHAOS-CT), 2.80 (LiTS), 2.64 (AMOS), 6.72 (CHAOS-MRI), and 3.66 (DSB). Even when compared to annotations supplemented by computational tools (second row), our framework maintained its advantage, achieving reduction factors of 2.33 (CHAOS), 1.30 (LiTS), 1.46 (AMOS), and 1.11 (CHAOS-MRI). These results validate the practical efficacy of our method in minimizing the manual effort required for high-quality annotations, thereby significantly enhancing the annotation process's efficiency.

Figure 5.1 further illustrate the robustness of our framework in terms of segmentation accuracy during iterative user interaction loops on the CHAOS and LiTS datasets. Our approach consistently outperformed state-of-the-art methods, achieving a Dice score exceeding 0.9 within an average of fewer than four user interaction loops. The observed performance highlights the superior adaptability of our framework, which leverages iterative user feedback and dynamic test-time parameter updates to refine segmentation quality progressively. Compared to baseline methods, our approach demonstrates unparalleled accuracy and efficiency, consistently achieving superior results across all interaction levels and evaluated datasets.

In addition, Figure 5.1 provides a detailed visualization of the improvement in Dice score with each user interaction, confirming that our framework attains higher accuracy with fewer interactions. These findings underscore the utility of our approach in clinical scenarios, where reducing manual annotation time while maintaining segmentation precision is critical.

Method		Time Taken for the Datasets (in min.)				
		CHAOS (84 CT Sl.)	LiTS (104 CT Sl.)	AMOS (243 CT Sl.)	CHAOS (70 MRI Sl.)	DSB (Cell Image)
Manual Annot. Time		63.00	84.00	140.00	57.00	11.00
Annot. Time$^{(prev.\,res)}$		32.00	39.00	79.00	39.00	-
GrabCut [90]	User Time	66.10	67.00	89.00	61.30	8.00
	Machine Time	30.00	26.00	36.00	22.00	5.00
f-BRS [40]	User Time	52.18	58.43	100.65	40.3	8.00
	Machine Time	5.35	6.34	15.00	9.52	3.00
RiTM [41]	User Time	48.82	54.31	110.75	42.30	8.00
	Machine Time	4.79	5.78	13.44	8.92	2.00
iSegFormer [37]	User Time	35.23	38.55	-	-	3.00
	Machine Time	5.21	4.66	-	-	1.32
PseudoClick [60]	User Time	30.44	35.45	67.22	40.11	3.00
	Machine Time	5.01	5.45	12.86	8.44	1.12
FocalClick [36]	User Time	25.33	31.50	-	-	4.00
	Machine Time	4.32	5.21	-	-	0.78
Our Method (UNet as Backbone)	User Time	9.24	24.41	40.23	36.72	3.00
	Machine Time	4.45	5.50	12.85	8.47	1.40

Table 5.1 For each method, we report the total time (in minutes) to reach a Dice Score of 0.95 for the CT Volume. Our method reaches the target with the least User Time among all the methods. Here, 'Manual Annot. Time' gives the time taken when the user performs the annotation from scratch on each slice. On the other hand 'Annot Time (prev. res)' give the time taken for the annotation when the annotator copies the annotation of the last slice and makes the necessary changes to fit it on the current slice. The term 'SL' is used to represent the number of slices. "84 CT SL" means that we have 84 Slices of the CT Modality.

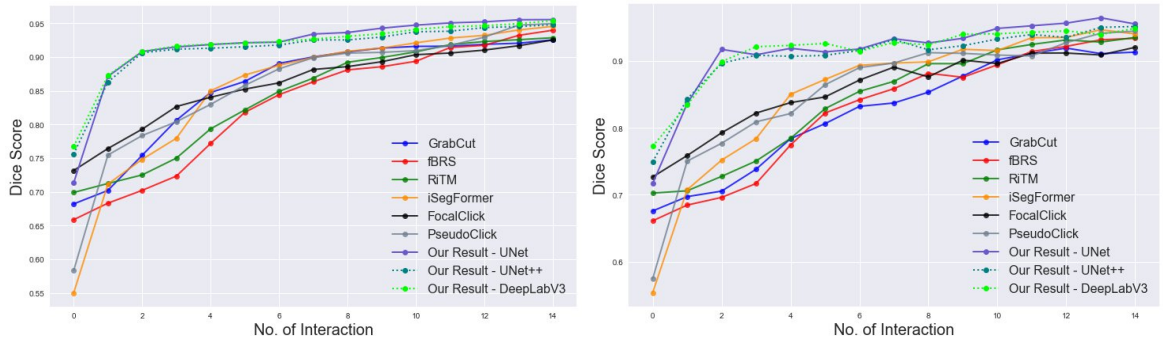

Figure 5.1 Improvement in Dice Score per user interaction: Our framework achieves the highest accuracy with minimum user interactions. Left: CHAOS (CT), Right: LiTS (CT)

5.1.2 Qualitative Results with Interactive Iterative Segmentation (IIS)

We provide a qualitative comparison of segmentation results across two consecutive slices using all the evaluated methods. These visual results effectively demonstrate the progressive refinement in segmentation quality achieved through successive interactions. Moreover, they underscore the superior performance of our proposed method, which produces segmentation outputs that exhibit better adherence to anatomical structures and finer details. This improved performance is particularly evident in the final segmentation results, where our approach achieves more accurate and visually coherent segmentations compared to state-of-the-art methods.

Figures 5.2, 5.5, 5.4 further illustrate the segmentation results on the LiTS, CHAOS, and AMOS datasets, respectively, showcasing comparisons across different methods, including GrabCut [90], fBRS [40], RiTM [41], iSegFormer [37], FocalClick [36], and our proposed approach. The results consistently highlight the ability of our method to achieve precise segmentations within 2–3 interactions, a finding that aligns with the conclusions drawn from the quantitative analysis. These results not only validate the efficiency of our approach but also emphasize its capability to achieve high-quality segmentations with minimal interaction, outperforming existing state-of-the-art techniques across diverse datasets and anatomical structures.

5.2 Comparison with prompting foundation models

In this study, we systematically compare our interactive segmentation method against several established generalized approaches, with a particular focus on state-of-the-art baselines, including Segment Anything Model (SAM) [61], MedSAM [74], ScribblePrompt [28], and MIDeepSeg [91]. To ensure a rigorous evaluation, we primarily consider the ViT-h variant of SAM [61], originally trained on natural images. SAM is designed as a foundation model for segmentation and operates using a variety of input modalities, including bounding boxes, user clicks, and the logits from previous segmentation predictions.

MedSAM [74] extends the capabilities of SAM by fine-tuning a ViT-B variant specifically for biomedical image segmentation tasks. This adaptation incorporates bounding box prompts and is trained on an extensive dataset comprising 1.5 million biomedical image segmentation pairs, thereby enhancing its applicability to medical imaging domains. In contrast, MIDeepSeg [91] is an interactive segmentation framework tailored for unseen medical imaging tasks. It employs interior margin points (positive clicks) as an initialization step, subsequently cropping the image based on these points before utilizing a convolutional neural network (CNN) for initial segmentation predictions. In our study, we evaluate the pre-trained MIDeepSeg model, which was originally developed for placenta segmentation in T2-weighted MRI scans.

Furthermore, ScribblePrompt [28] introduces a multi-modal user guidance strategy for interactive medical image segmentation. It leverages a combination of input prompts, including positive and negative scribbles, user clicks, and bounding boxes, while also incorporating prior segmentation predictions

Figure 5.2 Comparison on two LiTS slices. Left to right: original images, initial mask, Results after certain interactions, final segmentation mask, and Ground Truth. We have shown the results for Slice t and $t+1$ (in this case, slices 55 and 56) for GrabCut, fBRS, iSegFormer, FocalClick, and Our Results for the consecutive slices.

Figure 5.3 Comparison on two CHAOS slices. Left to right: original images, initial mask, Results after certain interactions, final segmentation mask, and Ground Truth.

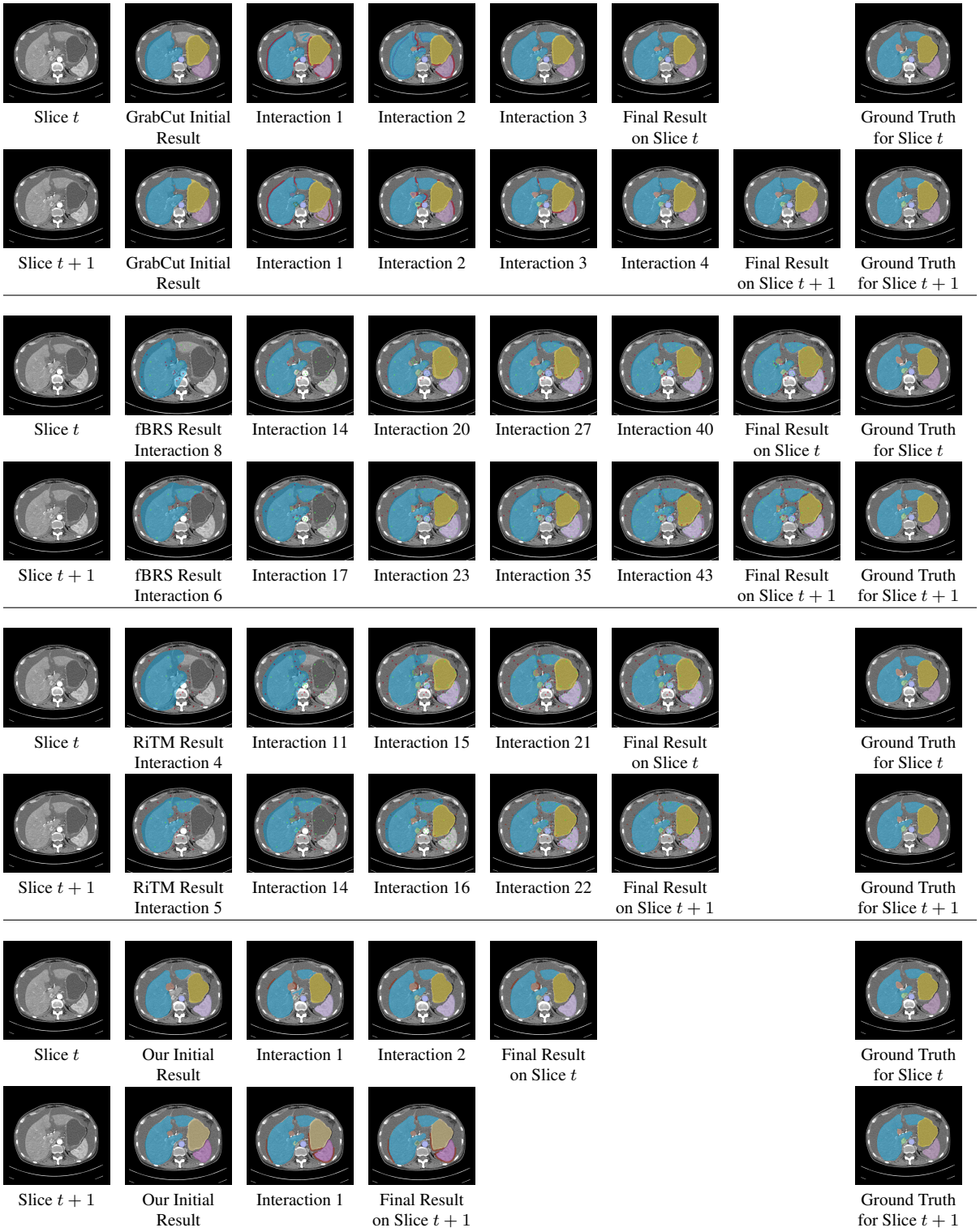

Figure 5.4 Comparison on two AMOS slices. Left to right: original images, initial mask, Results after certain interactions, final segmentation mask, and Ground Truth. We have shown the results for Slice t and $t+1$ (in this case, slices 175 and 176) for GrabCut, fBRS, RiTM, and Our Results for the consecutive slices.

to refine its outputs. By integrating multiple forms of user interaction, ScribblePrompt aims to enhance segmentation accuracy and adaptability across diverse medical imaging applications.

Through this comparative analysis, we aim to assess the efficacy of these existing interactive segmentation methods and highlight the advantages of our proposed approach in achieving accurate and efficient segmentation results across various medical imaging datasets.

5.2.1 Quantitative Results with prompt-based foundation models

Table 5.2 presents a comparative analysis of our proposed method against various foundational models utilizing different prompting strategies. The results demonstrate that our approach consistently outperforms these models, achieving Dice scores of 0.931, 0.936, and 0.942 on the CHAOS, LiTS, and Data Science Bowl (DSB) datasets, respectively. In contrast, ScribblePrompt attains Dice scores of 0.903, 0.913, and 0.929, while SAM records significantly lower scores of 0.741, 0.772, and 0.521. MedSAM, despite being fine-tuned for medical imaging, achieves moderate Dice scores of 0.804, 0.838, and 0.685 across the respective datasets.

Our analysis further reveals that foundational models exhibit performance stagnation after ten user interactions, producing suboptimal segmentation results. Specifically, we observed that click-based interactions in SAM and MedSAM can often misguide the model, resulting in segmentations that do not align with user expectations. This limitation stems from the fact that SAM was not originally designed to generalize well to click-based inputs, thereby reducing its adaptability to interactive segmentation tasks. While MedSAM demonstrates improved segmentation performance compared to other SAM-based models, its reliance on an initial bounding box as the primary input restricts its ability to incorporate negative click corrections effectively.

Methods	Dataset					
	CHAOS (CT)		LiTS (CT)		DSB (Cell)	
	Int. Dice	Final Dice	Int. Dice	Final Dice	Int. Dice	Final Dice
SAM [61]	0.628	0.741	0.573	0.772	0.498	0.521
MedSAM [74]	0.755	0.804	0.786	0.838	0.667	0.685
MIDeepSeg [91]		0.853	-	0.837	-	-
ScribblePrompt [28]	0.843	0.903	0.847	0.913	0.853	0.920
Our Method (UNet)	0.843	0.928	0.854	0.934	0.887	0.929
Our Method (UNet++)	0.881	0.933	0.898	0.912	0.867	0.931
Our Method (DeepLab-V3)	0.889	0.931	0.891	0.936	0.881	0.942

Table 5.2 User Interaction. We have effectively determined the mean Dice Score for segmentations predicted by different methods after ten rounds of user interaction. During each interaction loop, users can provide multiple scribbles, bounding boxes, and clicks to facilitate the segmentation process. We terminate the interaction loop if the method attains a dice score of 0.9.

In contrast, scribble-based interactions offer a more intuitive and flexible form of user guidance, enabling the segmentation algorithm to better comprehend the desired region of interest. The ability to integrate both positive and negative scribbles provides a more refined and user-adaptive segmenta-

tion process, ultimately contributing to the superior performance of our method across diverse medical imaging datasets.

5.2.2 Qualitative Results with Prompt-Based Foundation Models

Prompt-based foundation models, such as Segment Anything Model (SAM) and its medical imaging variant, MedSAM, have demonstrated significant advancements in zero-shot and few-shot segmentation tasks. However, our qualitative analysis reveals key limitations when these models are deployed for interactive medical image segmentation, particularly in a continual test-time adaptation setting.

Our observations indicate that while foundation models can initially provide reasonable segmentation outputs, their performance deteriorates with successive user interactions. After approximately ten user interactions, we observed stagnation in segmentation quality, with minimal improvements despite additional prompts. This suggests that these models lack the ability to effectively integrate new user feedback over time, which is crucial for refining medical image segmentations dynamically.

A key limitation we encountered was the sensitivity of foundation models to click-based prompts. Our experiments with SAM and MedSAM demonstrated that click inputs, especially in cases requiring fine-grained corrections, often misguide the model, leading to incorrect segmentation outputs. This is particularly evident in cases where the model fails to differentiate between foreground and background structures based on sparse user interactions. SAM, in particular, struggles with generalizing to click-based interactions, as it was not explicitly trained to handle such inputs. MedSAM, while showing better performance than standard SAM in medical imaging tasks, remains constrained by its reliance on the initial bounding box as its primary input. Notably, MedSAM does not effectively utilize negative clicks, limiting its flexibility in refining segmentation boundaries.

In contrast, we found that scribble-based interactions provide a more intuitive and effective way for users to communicate their desired segmentation adjustments. Scribbles allow users to delineate complex structures more accurately, and our continual test-time adaptation framework leverages these inputs to progressively refine the segmentation over multiple slices. Unlike conventional foundation models that operate in a static inference mode, our approach enables continual learning from user corrections, ensuring improved segmentation accuracy and consistency across volumetric data.

The advantages of our continual test-time training framework become evident when evaluating its ability to retain knowledge from past user interactions while adapting to new slices. Unlike SAM and MedSAM, which process each slice independently without maintaining inter-slice consistency, our teacher-student paradigm ensures that segmentation refinements persist across the volume. The teacher model aggregates knowledge across slices via an exponential moving average (EMA), preserving anatomical coherence, while the student model dynamically adapts to the current user input. This mitigates the common issue of segmentation drift observed in foundation models, where incorrect corrections in one slice do not carry forward effectively to subsequent slices.

Figure 4 illustrates a comparative visualization of segmentation results using foundation models. The figure highlights how SAM and MedSAM struggle with interactive refinements beyond initial predic-

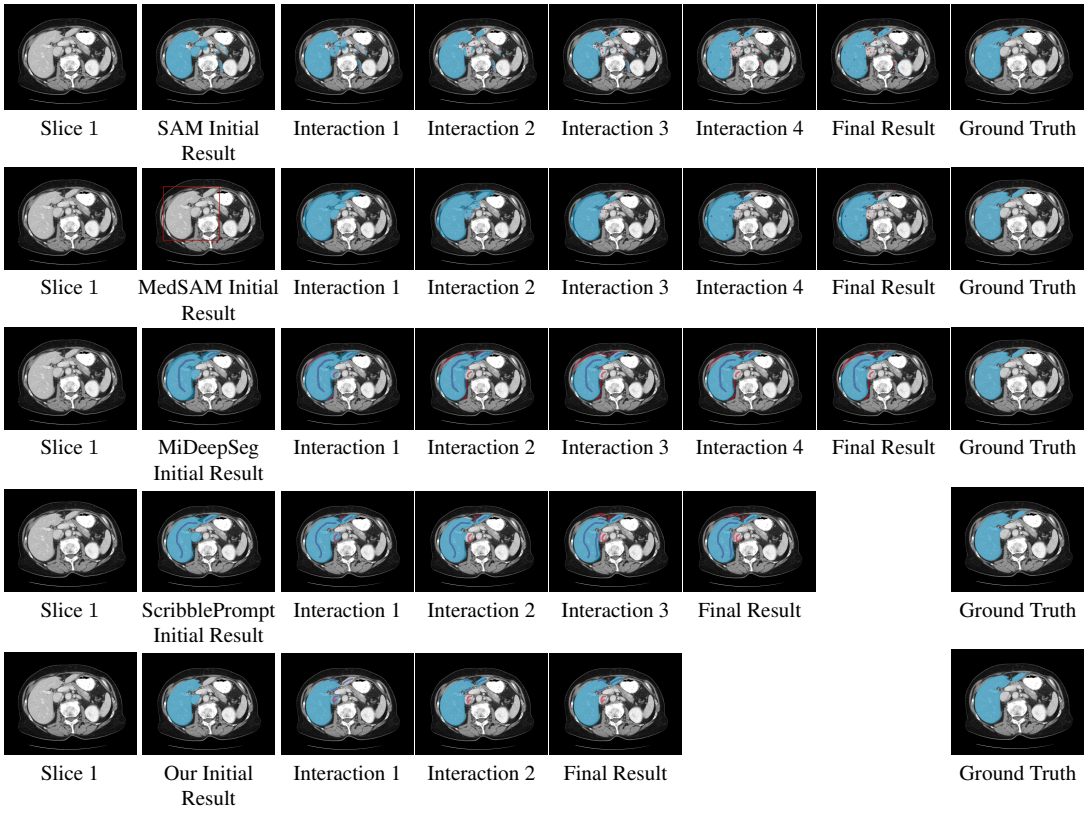

Figure 5.5 Comparison on LiTS slice for Foundational Models. Left to right: original images, initial mask, Results after each interaction, final segmentation mask, and Ground Truth. We have shown SAM, MedSAM, MiDeepSeg, ScribblePrompt and Our Result results. We have evaluated the same results for other datasets added in our supplementary work

tions, whereas our framework continues to refine segmentation quality with ongoing user interactions. These findings underscore the importance of integrating continual learning mechanisms into interactive medical image segmentation frameworks to overcome the limitations of existing prompt-based models.

Chapter 6

Conclusion

This thesis has introduced a novel framework for interactive medical image segmentation that leverages continual test-time training in a teacher-student learning paradigm. Addressing the limitations of existing static pre-trained models and prompt-based foundation models, our approach dynamically adapts to user interactions while maintaining inter-slice consistency. By integrating continual learning mechanisms into the segmentation process, we ensure that the model not only refines its predictions for the current slice but also retains and propagates relevant knowledge across the entire volumetric scan.

A key contribution of this work is the formulation of a test-time adaptation strategy that optimally balances short-term learning for immediate correction with long-term knowledge retention. The student model rapidly adapts to user scribbles and evolving segmentation needs, while the teacher model aggregates stable representations via an exponential moving average (EMA) update mechanism. This prevents overfitting to a single slice while preserving anatomical coherence across the volume. Extensive experiments on diverse medical imaging datasets, including CT, MRI, and microscopic cell images, demonstrate that our framework achieves superior segmentation accuracy with fewer user interactions compared to state-of-the-art interactive segmentation approaches.

The implications of this research extend beyond interactive segmentation. By demonstrating the effectiveness of continual test-time learning, this thesis provides a foundation for future advancements in real-time adaptation of deep learning models in medical imaging. The proposed approach can be further extended to other medical imaging tasks, including anomaly detection, multi-modal segmentation, and real-time clinical decision support. Future work could explore enhancing the framework with self-supervised learning techniques, integrating advanced uncertainty quantification methods, and expanding its applicability to large-scale medical datasets.

In conclusion, this thesis represents a significant step forward in interactive medical image segmentation by introducing a robust and adaptive learning framework. This work paves the way for more accurate, reliable, and user-friendly segmentation tools in clinical workflows by addressing the fundamental challenges of model adaptation, user interaction efficiency, and inter-slice consistency.

Bibliography

[1] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Computers in Biology and Medicine*, vol. 112, p. 103375, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482519302525

[2] J. Wahsner, E. M. Gale, A. Rodríguez-Rodríguez, and P. Caravan, "Chemistry of mri contrast agents: current challenges and new frontiers," *Chemical reviews*, vol. 119, no. 2, pp. 957–1057, 2018.

[3] P. J. Slomka, D. Dey, A. Sitek, M. Motwani, D. S. Berman, and G. Germano, "Cardiac imaging: working towards fully-automated machine analysis & interpretation," *Expert review of medical devices*, vol. 14, no. 3, pp. 197–212, 2017.

[4] W. C. Miller-Hance and R. Gertler, "Essentials of cardiology," in *A Practice of Anesthesia for Infants and Children*. Elsevier, 2019, pp. 355–392.

[5] R. Lattanzi, M. Viceconti, M. Petrone, P. Quadrani, and C. Zannoni, "Applications of 3d medical imaging in orthopaedic surgery: Introducing the hip-op system." in *3DPVT*, 2002, pp. 808–811.

[6] K. C. Wong, "3d-printed patient-specific applications in orthopedics," *Orthopedic research and reviews*, pp. 57–66, 2016.

[7] J. Weese and C. Lorenz, "Four challenges in medical image analysis from an industrial perspective," pp. 44–49, 2016.

[8] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on mri," *arXiv preprint arXiv:1811.10052*, 2018.

[9] B. Heim, F. Krismer, R. De Marzi, and K. Seppi, "Magnetic resonance imaging for the diagnosis of parkinson's disease," *Journal of neural transmission*, vol. 124, pp. 915–964, 2017.

[10] H. Liu, D. Hu, H. Li, and I. Oguz, "Medical image segmentation using deep learning," *Machine Learning for Brain Disorders*, pp. 391–434, 2023.

[11] P. Li, W. Wu, L. Liu, F. M. Serry, J. Wang, and H. Han, "Automatic brain tumor segmentation from multiparametric mri based on cascaded 3d u-net and 3d u-net++," *Biomedical Signal Processing and Control*, vol. 78, p. 103979, 2022.

[12] T. Ding, S. J. Wagner, A. H. Song, R. J. Chen, M. Y. Lu, A. Zhang, A. J. Vaidya, G. Jaume, M. Shaban, A. Kim *et al.*, "Multimodal whole slide foundation model for pathology," *arXiv preprint arXiv:2411.19666*, 2024.

[13] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep learning-enabled medical computer vision," *NPJ digital medicine*, vol. 4, no. 1, p. 5, 2021.

[14] X. Liu, L. Song, S. Liu, and Y. Zhang, "A review of deep-learning-based medical image segmentation methods," *Sustainability*, vol. 13, no. 3, p. 1224, 2021.

[15] A. Yala, P. G. Mikhael, F. Strand, G. Lin, K. Smith, Y.-L. Wan, L. Lamb, K. Hughes, C. Lehman, and R. Barzilay, "Toward robust mammography-based models for breast cancer risk," *Science Translational Medicine*, vol. 13, no. 578, p. eaba4373, 2021.

[16] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[17] K. Jamart, Z. Xiong, G. D. Maso Talou, M. K. Stiles, and J. Zhao, "Mini review: Deep learning for atrial segmentation from late gadolinium-enhanced mris," *Frontiers in Cardiovascular Medicine*, vol. 7, p. 86, 2020.

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

[19] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. 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*. Springer, 2016, pp. 424–432.

[20] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

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

[22] A. Popovic, M. De la Fuente, M. Engelhardt, and K. Radermacher, "Statistical validation metric for accuracy assessment in medical image segmentation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 2, pp. 169–181, 2007.

[23] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET image processing*, vol. 16, no. 5, pp. 1243–1267, 2022.

[24] X. Luo, G. Wang, T. Song, J. Zhang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "Mideepseg: Minimally interactive segmentation of unseen objects from medical images using deep learning," *Medical image analysis*, vol. 72, p. 102102, 2021.

[25] K. A. Philbrick, A. D. Weston, Z. Akkus, T. L. Kline, P. Korfiatis, T. Sakinis, P. Kostandy, A. Boonrod, A. Zeinoddini, N. Takahashi *et al.*, "Ril-contour: a medical imaging dataset annotation tool for and with deep learning," *Journal of digital imaging*, vol. 32, pp. 571–581, 2019.

[26] T. Sakinis, F. Milletari, H. Roth, P. Korfiatis, P. Kostandy, K. Philbrick, Z. Akkus, Z. Xu, D. Xu, and B. J. Erickson, "Interactive segmentation of medical images through fully convolutional neural networks," *arXiv preprint arXiv:1903.08205*, 2019.

[27] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin *et al.*, "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE transactions on medical imaging*, vol. 37, no. 7, pp. 1562–1573, 2018.

[28] H. E. Wong, M. Rakic, J. Guttag, and A. V. Dalca, "Scribbleprompt: Fast and flexible interactive segmentation for any medical image," *arXiv preprint arXiv:2312.07381*, 2023.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[30] R. Benenson, S. Popov, and V. Ferrari, "Large-scale interactive object segmentation with human annotators," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 700–11 709.

[31] D. Acuna, H. Ling, A. Kar, and S. Fidler, "Efficient interactive annotation of segmentation datasets with polygon-rnn++," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 859–868.

[32] N. Kim, B. Kang, and Y. Cho, "Split-gcn: Effective interactive annotation for segmentation of disconnected instance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 9256–9263, 2022.

[33] H. Ling, J. Gao, A. Kar, W. Chen, and S. Fidler, "Fast interactive object annotation with curve-gcn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5257–5266.

[34] Z. Lin, Z. Zhang, L.-Z. Chen, M.-M. Cheng, and S.-P. Lu, "Interactive image segmentation with first click attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 339–13 348.

[35] X. Chen, Z. Zhao, F. Yu, Y. Zhang, and M. Duan, "Conditional diffusion for interactive segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7345–7354.

[36] X. Chen, Z. Zhao, Y. Zhang, M. Duan, D. Qi, and H. Zhao, "Focalclick: Towards practical interactive image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1300–1309.

[37] Q. Liu, Z. Xu, Y. Jiao, and M. Niethammer, "isegformer: interactive segmentation via transformers with application to 3d knee mr images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 464–474.

[38] Z. Lin, Z.-P. Duan, Z. Zhang, C.-L. Guo, and M.-M. Cheng, "Focuscut: Diving into a focus view in interactive segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2637–2646.

[39] Y. Hao, Y. Liu, Z. Wu, L. Han, Y. Chen, G. Chen, L. Chu, S. Tang, Z. Yu, Z. Chen *et al.*, "Edgeflow: Achieving practical interactive segmentation with edge-guided flow," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1551–1560.

[40] K. Sofiiuk, I. Petrov, O. Barinova, and A. Konushin, "f-brs: Rethinking backpropagating refinement for interactive segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8623–8632.

[41] K. Sofiiuk, I. A. Petrov, and A. Konushin, "Reviving iterative training with mask guidance for interactive segmentation," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 3141–3145.

[42] B. Sambaturu, A. Gupta, C. Jawahar, and C. Arora, "Efficient and generic interactive segmentation framework to correct mispredictions during clinical evaluation of medical images," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*. Springer, 2021, pp. 625–635.

[43] X. Chen, Y. S. J. Cheung, S.-N. Lim, and H. Zhao, "Scribbleseg: Scribble-based interactive image segmentation," *arXiv preprint arXiv:2303.11320*, 2023.

[44] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool, "Deep extreme cut: From extreme points to object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 616–625.

[45] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[46] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[47] T. Kontogianni, M. Gygli, J. Uijlings, and V. Ferrari, "Continuous adaptation for interactive object segmentation by learning from corrections," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, 2020, pp. 579–596.

[48] Y. Hao, Y. Liu, J. Peng, H. Xiong, G. Chen, S. Tang, Z. Chen, and B. Lai, "Rais: Robust and accurate interactive segmentation via continual learning," *arXiv preprint arXiv:2210.10984*, 2022.

[49] A. Elangovan and T. Jeyaseelan, "Medical imaging modalities: a survey," in *2016 International Conference on emerging trends in engineering, technology and science (ICETETS)*. ieee, 2016, pp. 1–4.

[50] M. Mazonakis and J. Damilakis, "Computed tomography: What and how does it measure?" *European journal of radiology*, vol. 85, no. 8, pp. 1499–1504, 2016.

[51] R.-J. M. Van Geuns, P. A. Wielopolski, H. G. de Bruin, B. J. Rensing, P. M. Van Ooijen, M. Hulshoff, M. Oudkerk, and P. J. de Feyter, "Basic principles of magnetic resonance imaging," *Progress in cardiovascular diseases*, vol. 42, no. 2, pp. 149–156, 1999.

[52] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020.

[53] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[54] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of digital imaging*, vol. 32, pp. 582–596, 2019.

[55] S. M. Khaniabadi, H. Ibrahim, I. A. Huqqani, F. M. Khaniabadi, H. A. M. Sakim, and S. S. Teoh, "Comparative review on traditional and deep learning methods for medical image segmentation," in *2023 IEEE 14th control and system graduate research colloquium (ICSGRC)*. IEEE, 2023, pp. 45–50.

[56] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, no. 1, pp. 221–248, 2017.

[57] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. 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*. Springer, 2018, pp. 3–11.

[58] J. Zhang, Y. Xie, Y. Wang, and Y. Xia, "Inter-slice context residual learning for 3d medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 2, pp. 661–672, 2020.

[59] S. Zhang, J. H. Liew, Y. Wei, S. Wei, and Y. Zhao, "Interactive object segmentation with inside-outside guidance," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 234–12 244.

[60] Q. Liu, M. Zheng, B. Planche, S. Karanam, T. Chen, M. Niethammer, and Z. Wu, "Pseudoclick: Interactive image segmentation with click imitation," in *European Conference on Computer Vision*. Springer, 2022, pp. 728–745.

[61] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[62] H. Wang, P. K. A. Vasu, F. Faghri, R. Vemulapalli, M. Farajtabar, S. Mehta, M. Rastegari, O. Tuzel, and H. Pouransari, "Sam-clip: Merging vision foundation models towards semantic and spatial understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3635–3647.

[63] J. Cheng, J. Ye, Z. Deng, J. Chen, T. Li, H. Wang, Y. Su, Z. Huang, J. Chen, L. Jiang *et al.*, "Sam-med2d," *arXiv preprint arXiv:2308.16184*, 2023.

[64] V. I. Butoi, J. J. G. Ortiz, T. Ma, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Universeg: Universal medical image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 438–21 451.

[65] S. Czolbe and A. V. Dalca, "Neuralizer: General neuroimage analysis without re-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6217–6230.

[66] X. Wang, W. Wang, Y. Cao, C. Shen, and T. Huang, "Images speak in images: A generalist painter for in-context visual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6830–6839.

[67] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," *Advances in neural information processing systems*, vol. 32, 2019.

[68] S. He, R. Bao, J. Li, J. Stout, A. Bjornerud, P. E. Grant, and Y. Ou, "Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets," *arXiv preprint arXiv:2304.09324*, 2023.

[69] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: an experimental study," *Medical Image Analysis*, vol. 89, p. 102918, 2023.

[70] Y. Zhang, Z. Shen, and R. Jiao, "Segment anything model for medical image segmentation: Current applications and future directions," *Computers in Biology and Medicine*, p. 108238, 2024.

[71] S. Mohapatra, A. Gosai, and G. Schlaug, "Sam vs bet: A comparative study for brain extraction and segmentation of magnetic resonance images using deep learning," *arXiv preprint arXiv:2304.04738*, 2023.

[72] P. Zhang and Y. Wang, "Segment anything model for brain tumor segmentation," *arXiv preprint arXiv:2309.08434*, 2023.

[73] S. Roy, T. Wald, G. Koehler, M. R. Rokuss, N. Disch, J. Holzschuh, D. Zimmerer, and K. H. Maier-Hein, "Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model," *arXiv preprint arXiv:2304.05396*, 2023.

[74] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[75] T. Kitrungrotsakul, I. Yutaro, L. Lin, R. Tong, J. Li, and Y.-W. Chen, "Interactive deep refinement network for medical image segmentation," *arXiv preprint arXiv:2006.15320*, 2020.

[76] J. Sun, Y. Shi, Y. Gao, L. Wang, L. Zhou, W. Yang, and D. Shen, "Interactive medical image segmentation via point-based interaction and sequential patch learning," *arXiv preprint arXiv:1804.10481*, 2018.

[77] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv preprint arXiv:1612.02649*, 2016.

[78] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *International conference on machine learning*. PMLR, 2020, pp. 6028–6039.

[79] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7472–7481.

[80] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2517–2526.

[81] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *International conference on machine learning*. PMLR, 2020, pp. 9229–9248.

[82] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," *arXiv preprint arXiv:2006.10726*, 2020.

[83] A. Douillard, Y. Chen, A. Dapogny, and M. Cord, "Plop: Learning without forgetting for continual semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4040–4050.

[84] Q. Wang, O. Fink, L. Van Gool, and D. Dai, "Continual test-time domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7201–7211.

[85] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in neural information processing systems*, vol. 31, 2018.

[86] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv preprint arXiv:1706.02690*, 2017.

[87] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan *et al.*, "Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69, p. 101950, 2021.

[88] P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, G. E. H. Mamani, G. Chartrand *et al.*, "The liver tumor segmentation benchmark (lits)," *Medical Image Analysis*, vol. 84, p. 102680, 2023.

[89] Y. Ji, H. Bai, C. Ge, J. Yang, Y. Zhu, R. Zhang, Z. Li, L. Zhanng, W. Ma, X. Wan *et al.*, "Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation," *Advances in neural information processing systems*, vol. 35, pp. 36 722–36 732, 2022.

[90] C. Rother, V. Kolmogorov, and A. Blake, "" grabcut" interactive foreground extraction using iterated graph cuts," *ACM transactions on graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.

[91] X. Luo, G. Wang, T. Song, J. Zhang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "Mideepseg: Minimally interactive segmentation of unseen objects from medical images using deep learning," *Medical image analysis*, vol. 72, p. 102102, 2021.